STATISTICAL GENETICS AND MOLECULAR EVOLUTION OF MAJOR HISTOCOMPATIBILITY COMPLEX GENES

A thesis

submitted in partial fulfilment

of the requirements for the Degree

of

Doctor of Philosophy

in

Molecular and Cellular Biology

at the

University of Canterbury

by

David H. Bos

University of Canterbury

2005

Table of Contents

Table of Figures

Table of Tables

ABSTRACT

MHC region genes have been the subject of molecular evolutionary studies both from single species and from a variety of taxa. The African clawed frog, *Xenopus laevis,* provides a good model for the study of immune genes such as the MHC class Ia because of the genomic architecture of the MHC region. Herein, I investigate 1) the molecular evolution of the MHC class Ia gene at the population level in *X. laevis,* and 2) the evolution of proteasome subunits *psmb5* and *lmp7* following duplication from their common ancestral locus using a phylogenetic sampling of mainly vertebrate taxa. Model-based maximum likelihood statistical procedures are used in an effort to overcome typical problems associated with complex patterns of molecular evolution at these loci.

In this thesis I present several new findings, and Chapters I and II focus on phylogenetic investigation of proteasome subunits. Results indicate that several evolutionary mechanisms operate on *lmp7* that makes phylogenetic reconstruction of this locus difficult. I show that analysis of this gene is sensitive to the particular assumptions of various models of nucleotide evolution commonly used for phylogenetics. I also investigate whether or not natural selection operated differentially on duplicates of the proto-*lmp7* gene locus. I provide evidence that positive Darwinian evolution contributed to the functional divergence of gene family members derived from this locus—making this one of the few examples of positive natural selection operating on a protein with housekeeping functions.

Several new and major findings are also presented for the *X. laevis* class Ia MHC gene in Chapters III, IV and V. For the first time I provide robust estimates of substitution rates that show the operation of natural selection on peptide binding region (PBR) amino acids of the class Ia gene. I also show for the first time that intralocus recombinations are a major source of variation in the class Ia gene in *X. laevis.* Patterns of polymorphism at the class Ia locus are investigated in greater detail, and provide evidence for a molecular basis driving the coevolution of functionally linked genes. Combining data from other species, my results also demonstrate that the mode of MHC class Ia evolution is different than the classical paradigm detailed in mammals. Finally, my research is the first to demonstrate that non-linkage of the class I and class II genes in a single genomic region is not always necessary for this mode of class Ia evolution, as previously expected.

GENERAL INTRODUCTION

HUMAN MHC GENOMICS

The Major Histocompatibility Complex (MHC) is a large genetic region of the human genome found on the short arm of chromosome 6. The MHC contains over 220 gene loci and spans 3.6 Mbp, making it the most gene-dense region of the human genome (Beck and Trowsdale 2000; The MHC Sequencing Consortium 1999). Perhaps not coincidentally, many of the gene loci in the MHC encode proteins that are part of the immune system. In humans the MHC is called Human Leukocyte Antigen (HLA), and the MHC is annotated differently in other taxa; here the term MHC is used universally for the sake of simplicity, although established nomenclature of MHC of various taxa is recognized. The MHC has historically been divided into three regions: the class II region is found near the centromere, the class III region and the class I region is nearest to the telomere.

The class II region contains the class II MHC genes, and all other loci in this region with known function is involved in the immune system (Beck and Trowsdale 1999; Beck and Trowsdale 2000). Within the class II region are located families of class II genes encoding proteins such as HLA-DP, -DQ, and -DR (Bontrop et al. 1999). This region also contains many older pseudogenes, yet this region is recently thought to have become more sensitive to expansion by duplication of loci. As a result of this constancy, some members of the class II MHC multigene family are evolutionarily stable and orthologous class II loci can be found in taxa from other mammalian orders (Hughes and Nei 1990). Interestingly, in humans the class II region contains *lmp2, lmp7, tap1* and *tap2*, which are genes that are involved in the processing of peptides for class I MHC proteins (Beck et al. 1992).

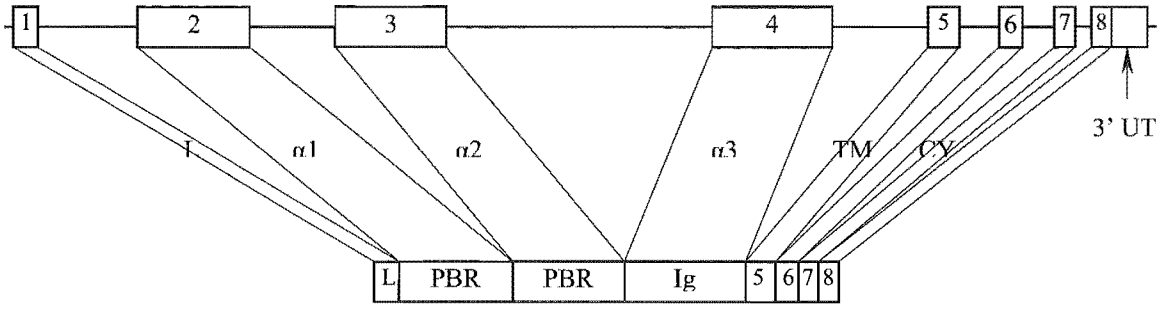
The class I region contains a multigene family of class I genes (some loci are designated *hla-a,-b*, and *-c*) and many other loci that are nonfunctional pseudogenes (Beck and Trowsdale 2000; Shiina et al. 1999b). Of the functional class I genes, some are designated as "classical" or "Ia" genes and others are "non-classical" or "Ib" loci. The distinction is somewhat subjective, but is largely based on characteristics such as patterns of expression, levels of polymorphism and conservation of certain amino acids. The human class I region is unlike the class II region because it is more prone to expansion through duplication and unequal crossing over. This leads to high

turnover of gene loci, and as a result orthologous copies of members of the class I multigene family can only be found in taxa from within the same order (Hughes and Nei 1989a; Yeager and Hughes 1999). Also a consequence of the high rates of tandem duplications is that the class I region has more pseudogenes than other regions (Beck and Trowsdale 2000). The high rate of duplications in the class I region is evident in the fact that various haplotypes in humans can differ in the number of gene loci. In contrast, the class III region appears to be more stable.

In humans, the class III region spans 800 Kbp, and is situated between the class I and the class II regions. This region undergoes very few duplications and as a result is almost totally devoid of pseudogenes, yet the class III region is the most gene-dense section of the MHC (Beck and Trowsdale 2000). While the class III region contains many non-immune genes, it is partially comprised of several loci involved in immune inflammation. Based on linkage of ancient genes, it is likely that the three regions of the MHC have different origins in primitive ancestors despite the fact that all three regions contain several genes involved in various immune functions. Parts of the class III region are thought to have the most ancient origin because copies of genes found in this region have been identified in syntenic relationships in invertebrate species (Beck and Trowsdale 2000). However, primary interest in the MHC region has focused on the class I and class II MHC molecules.

STRUCTURE AND FUNCTION OF MHC GENES

The class I molecules are heterodimeric proteins that are expressed in almost all cells. The gene for the α chain is encoded in the MHC and is comprised of eight exons that encode different domains of the protein (Figure 1). The α chain is associated with the β-2 microglobulin, which is encoded outside the MHC region. Exons 2 and 3 comprise the α1 and α2 domains respectively, which make up the peptide binding region (PBR) of the protein (Figures1 and 2). The crystal structure of the protein reveals that the α chain has three extracellular domains, a transmembrane domain, and an intracellular domain (Bjorkman et al. 1987b). β-2 microglobulin binds to the α3 domain of the protein to form an Ig-like C1 domain (Jones et al. 1998), and along with the PBR forms the extracellular portion of the molecule when it is expressed on the cell surface (Figure 2). The PBR is comprised of a base or floor which is made from a β-pleated sheet, and sides of the structure are made of α-helices (Bjorkman et al. 1987b).



MHC class I α chain

Figure 1. MHC class I α chain gene organization. Intron-exon structure of the α gene and corresponding domains of the protein are shown. Regions of the protein such as the PBR and Ig domain are also indicated.

The α-helices are spaced far enough apart to form a groove or channel between them, and the β-pleated sheet forms the bottom of the groove (Figure 2). The PBR is large enough to bind peptides of limited length: usually only peptides in the 7-9 amino acid length are found bound to class I MHC (Bjorkman et al. 1987a). The class II MHC also forms a PBR, but the structure is assembled differently.

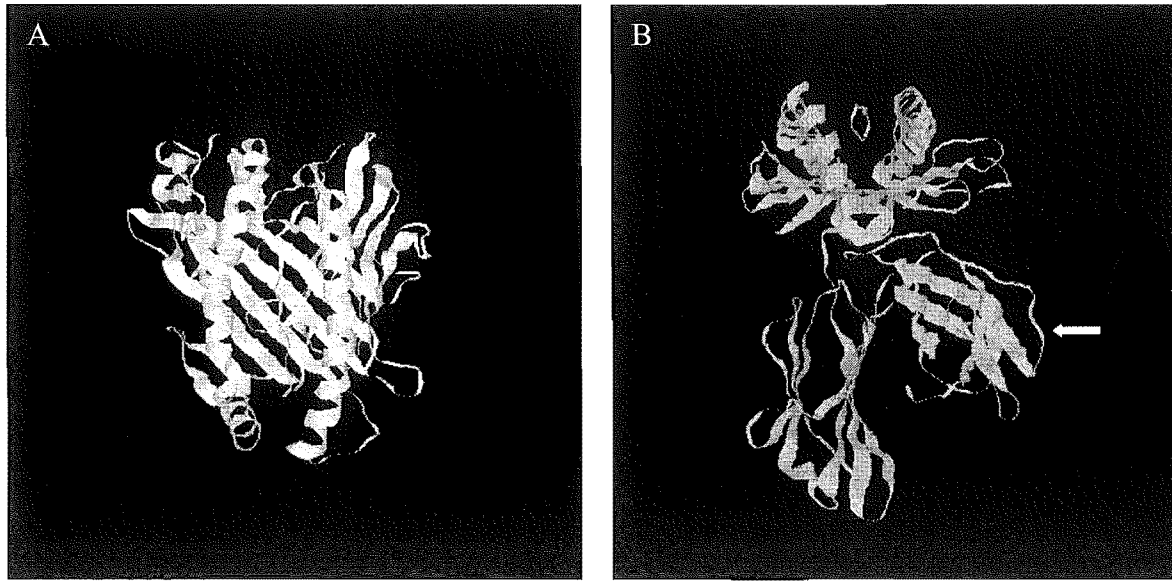


Figure 2. Three-dimensional structure of the MHC class I molecule. Extracellular domains of the protein are shown. A: the top view of the PBR is shown. The α helices and β-pleated sheet of the PBR are comprised of the α1 and α2 domains and can be distinguished. B: side view of the extracellular domains of the MHC class I molecule. The β2-microglobulin chain is indicated by the arrow. The groove of the PBR can be seen and is loaded with a small peptide.

Class II MHC are heterodimer proteins comprised of an α and β chain that are both encoded in the class II region of the MHC (Beck and Trowsdale 1999)(Figure 3). Unlike class I proteins where the PBR is encoded by a single chain of the heterodimer, each chain of the class II molecule makes up roughly half of the PBR and Ig-like C1 domain. The Ig extracellular structure of class II molecules is comprised of the α2 and β2 domains of the protein (Figure 4). The class II PBR is comprised of the α1 and β1 domains and structure of the PBR is similar to that of the class I PBR, with a β-pleated sheet and two α helices (Brown et al. 1993)(Figure 3). The class II PBR is more flexible than the class I, allowing peptides with a broader range of lengths to be bound. Peptides in the range of 11 to 17 amino acids in length

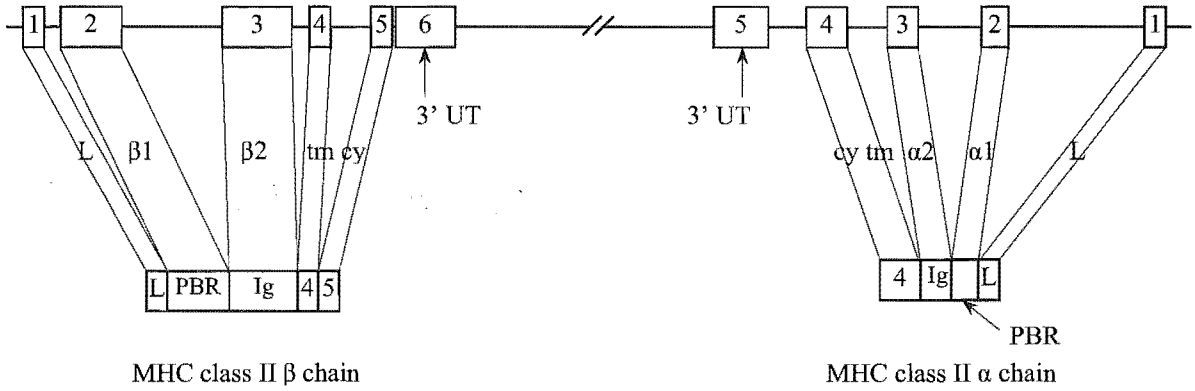


Figure 3. MHC class II α and β chain organization. Intron-exon structure of the genes are shown along with corresponding domains and structure of the complete class II protein.

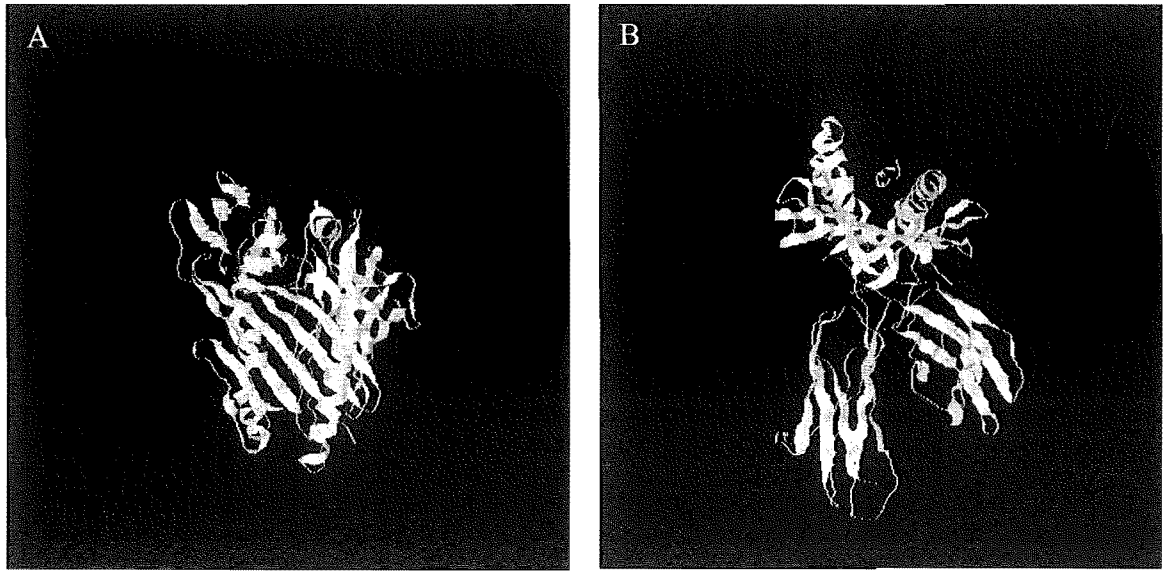


Figure 4. Three-dimensional structure of the class II protein. A: top view of the protein displays the α helices and β sheet of the PBR. B: Side view of the class II protein. The groove of the PBR with peptide is seen and Ig-like C1 domain is below the PBR.

are normally bound by class II molecules. The PBR of class I and class II MHC plays the primary role in the function of these proteins.

The main function of class I and class II molecules is to bind short peptides and present them to T cells (Abbas et al. 2000). Peptides bound by class I molecules are derived from within the cellular environment and are presented to CD^+ 8 T cells. Class II proteins bind peptides originating from extracellular proteins and are presented to CD^+ 4 helper T cells. T cells recognize the MHC bound peptides as either a normal component of the organism or a foreign/mutated protein fragment. When foreign peptides are discovered, then a cascade of immune signals is triggered and the infected cells are destroyed (Abbas et al. 2000). In many ways the MHC proteins are critical to the immune system, and they often provide cues for initial activation of the cellular immune response. Compared to the rest of the genome, MHC genes have many unusual features thought to have arisen due to intimate interaction with a wide variety of foreign pathogens.

POPULATION GENETICS OF MHC

MHC class I and class II genes are the most polymorphic loci in the human genome. The highest levels of diversity are concentrated in the PBR of the proteins, mainly in the α1 and α2 domains of class I and the β1 domain of class II (Figures 1 and 3). MHC polymorphism is also accompanied by high levels of allelic diversity in a population. There are over 200 known alleles for some MHC genes, and high levels of heterozygosity are found at class I and class II loci (Parham and Ohta 1996). Natural selection is thought to have promoted allelic diversity and polymorphism in the PBR, presumably as a result of selective pressure from the diverse array of foreign pathogens (Hughes and Nei 1988; Hughes and Nei 1989b). Interestingly, the selective forces have not brought any allele to fixation and there is no wild-type allele for the class I and class II MHC. Instead, many alleles in a population are found in nearly equal frequency. This is thought to be the result of balancing selection on the PBR rather than directional selection (Hughes and Yeager 1998). Balancing selection may also be the cause of other unusual features of MHC genetics.

Allelic lineages found in class I and class II loci may persist for very long periods of time. As mentioned above, orthologous copies of the human class I MHC can be found within the same order, and class II orthologues can be found in taxa from different orders. In a similar sense, allelic lineages of orthologous MHC loci

often predate the divergence of closely related species in which they are found (Figueroa et al. 1988; Lawlor et al. 1988; McConnell et al. 1988). This is particularly true of primate class II MHC (Bontrop et al. 1999), and the phenomenon has been termed "trans-species polymorphism". Evidence for trans-species polymorphism comes from phylogenetic studies showing that relationships among MHC alleles cluster by locus and MHC lineage, rather than by species. This is also true of class I loci, but due in part to the higher rate of duplication among these loci, the trans-species polymorphism is limited to more closely related species (Parham and Ohta 1996; Vogel et al. 1999). Although these phylogenetic patterns may have been caused by convergent evolution among MHC alleles from different species, the apparent lineage sharing among closely related species is thought to be primarily the result of polymorphism that predate some extant species (Yeager and Hughes 1999). The sustained persistence of allelic lineages beyond the formation of species is well beyond that expected for neutral alleles, but is consistent with theoretical expectations of allelic maintenance from balancing selection (Takahata and Nei 1990; Takahata et al. 1992). This trans-species polymorphism is an apparent paradox when considering the roles of recombination in MHC genes.

Evidence for recombination in playing a role in increasing the number of alleles at a locus has been found for some class I and class II genes (Jakobsen et al. 1998; Satta 1997; Takahata and Satta 1998). The class I *B* locus has been shown to have the highest rates of genetic exchange in humans (Hughes et al. 1993). The exchange events are thought to occur mostly among alleles at a locus rather than as interlocus events, although the latter may play a minor role in MHC evolution. In the class I *B* gene, most of the recombinations involve a very short "cassette" of 5-20 adjacent nucleotides, which are thought to occur through gene conversion (Jakobsen et al. 1998). There is also evidence that exon shuffling among alleles occurs, but this is a limited event compared to the smaller scale gene conversions, which are very prominent (Hughes et al. 1993). However, these exon shuffling events are very important because they often occur in intron II, which creates an allele with a new PBR (Figure 1). Class II genes may also under go recombinations, and the genetic exchange events may create an allele with a new PBR (Gyllensten et al. 1991). However, class II recombinations are rare and more limited than those of the class I *B*, as are other class I loci (Hughes et al. 1993).

MHC EVOLUTION

MAMMALS

Among all taxa, the MHC of humans is the most well characterized on the genomic, population, and allelic levels. The mode of MHC evolution was first ascertained from the study of mammalian models, which have shown that MHC loci evolve through a process of duplication, independent divergence and silencing (Gu and Nei 1999; Nei et al. 1997). Interlocus conversion and concerted evolution among gene family members seems to play a small role in MHC evolution (Hughes et al. 1993). Instead, individual loci diverge through point mutations, natural selection drives the fixation of many mutations in the PBR, and recombinations shuffle existing variation to create new alleles. Natural selection may also fix a proportion of recombinant alleles. Class I and class II loci differ in the rates and pattern duplication, divergence and silencing.

In the classical paradigm of MHC evolution, loci may differ in mode of evolution, but general trends among class I and class II loci are observed. Class I loci are more prone to duplications and have a higher rate of replacement while class II loci are generally more stable and long-lasting (Bontrop et al. 1999; Vogel et al. 1999). Class I loci have fewer orthologous relationships among related species when compared to class II genes (Yeager and Hughes 1999). Sharing of allelic lineages in class I genes of closely related species is also more uncommon when compared to class II allelic lineages, often due to the prominent role of recombination in some class I loci (Hughes et al. 1993). A clear pattern of recombination events is also not observed, probably because genetic exchanges are not localized but can occur almost anywhere along the gene and typically involve small stretches of nucleotides. Finally, polymorphism and diversity of mammalian MHC is multigenic, in that it is collective among loci of a gene family rather than concentrated in a single locus. While it has been known for a long time that different loci have different modes of evolution and there is flexibility in modes of evolution (Klein et al. 1993), the above basic relationships are well established and in the past have been implicitly assumed to be widespread. Only recently has the mode of MHC evolution been investigated in other taxa.

FISHES

Genomic organization of the MHC region in fishes has recently been the subject of study, and linkage patterns and genes are among the more adequately characterized

MHC. The structure and function of MHC class I and class II proteins are largely conserved since fishes and mammals last shared a common ancestor (Hashimoto et al. 1999). However, the organization of the MHC region in fishes is very different from that of mammals, in that the class I and class II regions of bony fishes are found in two separate regions of the genome (Bingulac-Popovic et al. 1997; Sato et al. 2000). No complete class III region has been detected, but multigene families of class I and class II loci define different localized regions of the MHC. The non-linkage of class I and class II regions in teleost fishes led to speculation that this organization might be the primitive, ancestral state of MHC linkage. This possibility was largely dispelled however, when it was established that the class I and class II regions are linked in sharks, the most primitive organism in which MHC proteins have been found (Ohta et al. 2000). Aside from common multigene families found in the class I region, in fishes this region is different from mammals because it contains genes involved in the processing of peptides for class I proteins (Ohta et al. 2002; Takami et al. 1997). The class I processing proteins encoded by LMP and TAP loci are found in the class II region of humans; however, they are found closely linked to class Ia genes in bony fishes (Graser et al. 1998; Michalova et al. 2000). Aside from differences in linkage pattern and locations of some genes, there are also distinctive evolutionary patterns of fish MHC genes (Figueroa et al. 2001; Shum et al. 2001).

Although few population studies on teleost fish MHC genes have been done, current evidence indicates that the mode of MHC evolution in fishes is different from the classical paradigm. MHC evolution in fishes follows the basic paradigm of evolution by gene duplication, independent divergence and natural selection (Miller et al. 2002). However, in fishes the pattern established by these forces has resulted in a different evolutionary trend for class I and class II genes (Figueroa et al. 2001; Shum et al. 2001). In salmonid fishes, the class I locus is subject to frequent recombinations, but they are unlike those of the mammals because they involve localized exon shuffling that is rare in mammals (Shum et al. 2001). Frequent interallelic exon shuffling results in a clear pattern of genetic exchange that is a fundamental feature of MHC evolution in some fishes. Exon shuffling events commonly create allelic diversity and are typically detected in intron II between the two exons that make up the PBR domain. Class Ia alleles in some fishes also have a higher level of polymorphism when compared to mammalian class Ia loci (Aoyagi et

al. 2002; Shum et al. 2001). MHC genes also differ in the rate of duplications and stability of loci when compared to mammals.

Fishes and mammals often differ markedly in patterns of polymorphism and age of allelic lineages among MHC gene families. Class I loci of fishes are stable, and as a result of their unchanging nature, allelic lineages are long-lasting and there is extensive trans-species polymorphism among bony fishes (Aoyagi et al. 2002; Figueroa et al. 2001; Shum et al. 2001). In contrast, the class II region is less stable in some species and more prone to duplications, disrupting trans-species polymorphisms more frequently. Consequently, allelic lineages in class II genes of some species are relatively short-lived and cluster in a species-specific manner compared to the intermingled clustering of class I alleles of different fish species (Shum et al. 2001). Other species of fishes display patterns consistent with stability and persistent lineages at both class I and class II loci (Figueroa et al. 2001). Also, in salmonid fishes the class Ia polymorphism is much higher than that of the class II loci, a pattern that is the opposite of that observed in mammals (Aoyagi et al. 2002; Shum et al. 2001). Another consequence of the stability of the class I region is that in fishes there is only one class Ia locus, but in mammals the class I region has several functional class Ia genes (Aoyagi et al. 2002; Shiina et al. 1999b; Shum et al. 2001). Therefore, mammalian MHC polymorphism is multigenic, whereas in salmonid fishes the polymorphism is multiallelic.

The differences observed between teleost fishes and humans have been attributed to the unique separation of class I and class II regions in fishes (Shum et al. 2001). In mammals, selection at a single locus affects all linked loci in the region. As a result, evolution at one locus is partially influenced by selection at another locus and evolution is compromised at both loci. These aspects of the genomic organization found in humans and other mammals may have led to the classical pattern of MHC evolution seen in mammals but not fishes. Instead, the MHC is divided into two unlinked regions of the bony fish genome, and this difference may have led to the differences in mode of evolution. Also, the larger intron sizes seen in some fishes may have influenced the evolution of MHC, especially aspects of genetic exchange. However, others have indicated that evolution of the different regions of the human MHC is unaffected by selective pressures at other MHC regions (O'hUigin et al. 2000), and that intron size has little effect on rates of exon shuffling of the MHC class I genes (Figueroa et al. 2001). The hypothesis of MHC region linkage and mode of

evolution has not been directly tested so it is unclear to what the vast differences in evolution between these two taxa can be attributed.

BIRDS

The genomic organization and evolution of the MHC in birds is also dissimilar from patterns seen in mammals. In chickens, there are two unlinked MHC regions, named the B complex and Rfp-Y. Polymorphic class I loci have been found in both regions, while class II loci are restricted to the B complex (Hess and Edwards 2002). Proteasome components found in the MHC of other taxa are not in the MHC of chickens, but other genes involved in processing of peptides for class I proteins, such as *tap1* and *tap2* are located in the vicinity of class I genes (Kaufman et al. 1999; Shiina et al. 1999a). Little work has been done to characterize the evolution of class I loci, but class II evolution has been studied in several taxa. When class II genes of birds are compared in a phylogenetic tree, different loci typically cluster by species (Edwards et al. 1995; Wittzell et al. 1999). This pattern differs from that seen in mammalian class II loci and indicates that class II loci have either been created by recent duplications or that divergence among loci has not been independent, but has been marked by homogenizing gene conversion. Both of these explanations differ from patterns of class II evolution in mammals, where loci originate from ancient duplications and diverge independently.

XENOPUS FROGS

MHC class I and class II proteins of the frog *Xenopus laevis* were among the first non-mammalian MHC to have been isolated and linkage patterns investigated. MHC proteins are similar to their mammalian counterparts in structure (Figure 1), with conserved residues crucial for maintaining the shape required for the function (Flajnik et al. 1991; Flajnik et al. 1984). *X. laevis* have a single MHC region, similar to humans, but unlike bony fishes (Nonaka et al. 1997a). In the MHC region, both class I and class II multigene families are found, as well as LMP and TAP loci, which are the class I processing protein genes. Like mammals and fishes, in *X. laevis* there are a number of class I non-classical (Ib) genes, but these are unlinked to the MHC region (Flajnik et al. 1993). This pattern is unusual since Ib loci are found in the MHC of mammals and fishes, typically in the class I region (Trowsdale 1995). Despite

similarities in the MHC region of *Xenopus* and humans, there are other noteworthy differences.

Unlike humans, the class I genes and the class I processing genes are tightly linked in *X. laevis* (Namikawa et al. 1995; Ohta et al. 1999). This pattern of linkage is common to both bony and cartilaginous fishes, indicating that the linkage pattern found in mammals is unusual. The linkage of class I genes and genes involved in the pathway of class I peptide processing seems more intuitive, since these proteins are functionally linked. In *X. laevis*, there is association between specific lineages of class I alleles and *lmp7* gene lineages (Nonaka et al. 2000). The association between specific lineages at functionally linked loci leads to speculation that specific alleles of *lmp7* coevolved with certain class I lineages for adaptive purposes. These coevolving allelic lineages have been maintained for many years in a single interbreeding species despite the possibility of recombination between them. The pattern of coevolution of class I processing genes and the class I genes themselves has also been observed in other taxa in which they are in close linkage (Kaufman 1999). Aside from differences in linkage pattern, there are other distinctions of the *X. laevis* MHC region.

In *X. laevis*, there is only a single class Ia locus, and at that locus only two alleles are expressed per individual even though *X. laevis* is a tetraploid species (Du Pasquier et al. 1977; Flajnik et al. 1999). The number of classical loci and expression pattern of class I genes is analogous to patterns in salmonid fishes. An unusual feature of *X. laevis* however, is that class I genes come in two forms that have different mRNA sizes. The difference is found at the intracellular domain of the sequences, which is extended in one lineage of alleles (Flajnik et al. 1999). Interestingly, alleles belonging to different lineages are as divergent as MHC class I genes from humans and mice. The polymorphism at the class I locus in 6 sequences from *X. laevis* and *X. laevis/X. gilli* hybrids exceeds the levels of polymorphism at the most polymorphic class I loci in humans, but is comparable to levels detected in salmonid fishes (Flajnik et al. 1999). The variation seen in *Xenopus* species makes it difficult to accurately estimate molecular evolutionary parameters such as substitution rates.

Class Ia loci are generally accepted as evolving under the evolutionary forces of natural selection. Evidence of an elevated nonsynonymous substitution rate usually provides strong evidence that balancing selection has occurred in the PBR (Hughes and Nei 1988). However, Flajnik (1999) was unable to demonstrate an elevated

nonsynonymous substitution rate in the class Ia gene of *Xenopus* using traditional methods. Although unable to provide strong evidence for natural selection, Flajnik et al. (1999) concluded that selection was probably a factor influencing MHC evolution in *Xenopus*, but that limitations of methods and the high level of diversity in the sequences prevented accurate estimation of substitution rates.

X. laevis class II genes have also been sequenced and characterized (Sato et al. 1993). There appears to be three classical class II loci in frogs and their structure and function and is also similar to humans. One exception is that intron sizes are much larger than humans and is more reminiscent of salmonid fishes (Kobari et al. 1995). Only a few sequences are available for class II sequences, so the level of polymorphism at a locus or other factors of interest are largely unknown.

In many aspects the organization of the MHC in *Xenopus* frogs is analogous to that found in salmonids. The tight linkage of class Ia and class I processing genes is similar in these species, and is unlike the pattern seen in humans, where the class I processing genes are found in the class II region. Despite many similarities in the linkage patterns of *X. laevis* MHC and that of salmonid fishes, one main difference is that there is only a single MHC region in *X. laevis*. While much is known of *X. laevis* MHC, many aspects of evolution remain to be investigated, namely whether or not the mode of evolution is more like the classical paradigm or like that found in salmonid fishes. In addition, one important aspect of the origin and evolution of the MHC region in vertebrates is the evolution of proteasome components that are involved in processing of MHC class I peptides.

PROTEASOME COMPONENTS OF THE MHC

Aside from MHC genes of the MHC region, the proteasome components found in the MHC are of particular interest and they play a crucial role in the functioning of MHC proteins. The LMP proteins are part of the 20S proteasome which is a vital housekeeping component of the cell. It is responsible for degradation of most cellular proteins, including regulatory proteins, transcription factors, catalytic enzymes and mutated or misfolded proteins (Rock et al. 1994). The 20S proteasome is made up of several components or subunits, and in vertebrates there are two types of subunits, the alpha and beta subunits (Coux 1996; Groll et al. 1997; Lowe et al. 1995). The 20S proteasome is comprised of seven subunits formed into a ring, and four rings are layered on top of one another to form a hollow cylinder. In vertebrates, each ring

comprises seven different α or β type subunits, and the proteasome consists of two α and two β type rings, with α subunit rings on each end. Three of the β subunits in each ring contain the active site of proteolysis, which is found inside the central channel of the proteasome (Arendt and Hochstrasser 1997; Heinemeyer et al. 1997; Seemuller et al. 1995). The ends of the 20S proteasome can be "capped" with the PA700 regulator (to form the 26S proteasome), the PA28 regulator, or one of each. These accessory protein complexes assist by regulating entry into the proteolytic core of the proteasome and by modifying the structure of the proteins destined for degradation (Gray et al. 1994; Groetterup et al. 1996; Tanaka and Kasahara 1998; Tanaka et al. 2000).

The products of proteasomal degradation are short peptides 3 to 30 residues in length, with the large majority of peptides at the shorter end of the size spectrum (Kisselev et al. 1999). These peptide products are typically destined for additional lysis into amino acids by cytoplasmic peptidases, although some are transported to cellular organelles. The action of the protease typically cleaves proteins after acidic, basic or hydrophobic amino acid residues through the action of the three active sites containing β subunits (Arendt and Hochstrasser 1997; Heinemeyer et al. 1997). However, stimulation of cells by gamma-interferon changes the biochemical profile of cleavage sites and size spectrum of peptide products generated by the proteasome (Boes et al. 1994; Driscoll et al. 1993). These changes are the result of replacement of the three conventionally expressed active β subunits by closely related gene family members that have been found only in the genomes of gnathostomes (Kandil et al. 1996). Expression of the three extra subunits is controlled by interferon-γ (which down-regulates their replaced components), and they are incorporated into newly assembled proteasomes (Akiyama et al. 1994; Gaczynska et al. 1993; Gaczynska et al. 1994).

Due to the exchange of β subunits that have the active site of proteolysis, the newly assembled proteasomes are functionally distinct from proteasomes with primarily housekeeping functions. The conventionally expressed subunits of the housekeeping proteasome that are exchangeable are called X, Y and Z (human genome database coded as PSMB5, PSMB6 and PSMB7 respectively) and they are replaced by LMP7 (PSMB8), LMP2 (PSMB9) and MECL1 (PSMB10) respectively (Coux 1996). When these three new subunits are in place, the action of the proteasome changes so that proteins are cleaved more frequently after hydrophobic

residues and less after acidic residues (Gaczynska et al. 1994). These peptides are then more frequently transported to the endoplasmic reticulum (ER) instead of being further digested by peptidases.

Transport to the ER occurs via the TAP transporter, comprised of members of the ATP-binding cassette (ABC) superfamily, which preferentially transports peptides with hydrophobic ends (Monaco and Nandi 1995). In the ER, these peptides are loaded onto MHC class I proteins and are displayed on the cell surface to cytotoxic T-cells (Coux 1996; Goldberg and Rock 1992; Rock et al. 2002; Tanaka and Kasahara 1998). Normal mechanisms that monitor the cellular environment for foreign pathogens or mutations use MHC genes and the conventionally expressed housekeeping proteasomes and proteolytic pathway (Rock et al. 1994), but this surveillance is enhanced by the alternative γ-interferon induced replacements in proteasome subunits and assembly of functionally distinct proteasomes.

Because proteasome components LMP7 and LMP2 are encoded in the MHC region, which also contains the class I genes and components of the TAP transporter (Flajnik and Kasahara 2001), and because of their primarily immunological function, proteasomes with these replacement subunits are called immunoproteasomes. Immunoproteasomes are functionally distinct from the conventionally expressed proteasomes due to the incorporation of active β subunits LMP2, LMP7 and MECL1. These proteins are similar in structure to their respective housekeeping counterparts (Brown et al. 1991; Glynne et al. 1991), and homology has been inferred based on this similarity. Linkage patterns and inferred homology have indicated that the loci encoding these proteins were possibly created by simultaneous chromosomal duplication of the more ancient *psmb5, 6* and *7* (Clark et al. 2000; Kasahara et al. 1996); however, the chromosomal duplication theory has been disputed.

Since homology is an evolutionary concept (Thornton and DeSalle 2000), Hughes (1997) used phylogenetic methods to establish homologous relationships among proteasome components. He tested the chromosomal duplication theory and used a molecular clock to estimate the timing of duplication events and some evidence that conflicts the chromosomal duplication theory. Despite investigations of these duplication events, the molecular evolutionary forces responsible for functional divergence since gene duplication have not been established, and conflict over differing evidence on some gene duplications has not been resolved. The conflict remains in part because duplication events in proteasome components took place

anciently, making it difficult to quantify parameters accurately and infer molecular evolutionary forces responsible for divergence since duplication.

AIMS OF THIS STUDY

Previous research on genes found in the MHC region has been hampered by difficulties in accurately estimating molecular evolutionary parameters and consequently making inference on the evolution of the MHC. For example, disagreement exists over the timing of the duplication of proteasome components and their subsequent evolution since duplication (Hughes 1997; Kasahara et al. 1996). Likewise, the complexity of factors acting on MHC genes such as balancing selection, recombination, duplications, and substitution accumulation have made it difficult to infer their evolutionary history (Hughes 1998; Hughes 2000; Kasahara 1999). As more data become available and improved methods for estimating evolutionary parameters are formulated, our understanding of the structure, diversity and function of DNA sequence will improve.

In this study I apply existing statistical methods that rely on mathematical models of DNA evolution to quantify molecular evolutionary parameters of the MHC class Ia and *lmp7* genes. Use of statistical- and phylo-genetic methods was prompted by the complexity of analysis and the unsatisfactory limitations of traditional pair-wise methods. Methods applied here typically involve the use of maximum likelihood (ML) statistics to optimize parameters such as phylogenetic branch lengths and substitution rates. Like most statistical methodologies, ML statistics relies on a mathematical model to approximate real processes, and in this case models approximate the process of molecular evolution.

In the molecular evolutionary context, models approximate factors that drive the changes (substitutions) in DNA sequences (Lio and Goldman 1998). These models can either use the individual nucleotide, or a codon triplet as the basic unit of evolution. These models often take into account the variation in substitution rates across different codons or nucleotides, the bias that prefers substitutions to nucleotides or codons that are similar in properties and the frequency of codons or nucleotides (Swofford et al. 1996). Models can also take into account changes in evolutionary processes in different evolutionary time periods (Yang and Nielsen 2002). While current models approximate many aspects of molecular evolution, they are considerable simplifications of real processes.

The evolution of DNA is highly complex, and many forces influence molecular evolution through mechanisms such as recombination and mutation (Hillis et al. 1996; Li 1997). Furthermore, population parameters such as mating system, size and growth affect changes in DNA. Many evolutionary and population parameters are dynamic and may operate for only a short time or affect different parts of a gene or genome in different ways. Essentially, evolutionary genetic elements and their interactions are so complex they can only be considered with a nearly infinite-order of complexity. However, in most cases one could imagine that there are a few major factors that affect evolution, many other factors that play a minor role, and many more that have a minuscule effect. In contrast to the high-orders complexity of molecular evolution, any data set will be finite in nature and thus have a limited amount of information on underlying evolutionary processes. When using statistical models, the objective is to approximate information (on underlying evolutionary processes) in the data, rather than to replicate the entire molecular evolutionary process.

Models that are highly complex compared to the finite data at hand are of limited value (Anderson et al. 2001). The most useful models maximize the amount and kind of information found in the data, typically major elements of evolution that are common to all samples of data. Use of a highly accurate model would result in highly precise estimates of parameters, but each would have a very large error due to the complexity of the model compared to the sample size of the data (Burnham and Anderson 2002). Furthermore, the overly-complex model may describe spurious processes found in the sample data set but not common to the larger population, and thus would be of limited value in predicting or making inference to the population as a whole (Anderson et al. 2001). Ideally, a model should accurately approximate the maximum amount of information in the data as simply as possible. Selecting the most useful model for a data set can be done objectively using practical information-theory statistical procedures (Burnham and Anderson 2002).

Using the ML framework, the objective of the thesis is to apply statistical genetic analysis to the evolution of MHC class Ia in *X. laevis* and *lmp7/psmb5* genes from among vertebrate taxa. This objective is motivated from differences seen among taxa in patterns of MHC evolution, and because of insufficiencies of traditional approaches in analyzing highly polymorphic sequences. These data represent two different types: one data set consists largely of population data from *X. laevis,* and the other is phylogenetic in nature and consists of *lmp7* and *psmb5* sequences from a

variety of taxa. Different evolutionary factors operate at the population and phylogenetic levels, and application of model-based statistical genetics on different types of data demonstrates flexibility in use and ability to characterize molecular evolution at different levels. The specific aims involving the use of these data are to:

- Investigate the mode of MHC evolution in *X. laevis*
- Characterize the evolutionary divergence of *psmb5* and *lmp7*
- Characterize the role of selection and inter-allelic recombination in alleles of the MHC class Ia locus of *X. laevis*
- Investigate the effects of models and objective model selection on the estimation of substitution rates in highly polymorphic MHC class Ia genes
- Investigate the effects of models and objective model selection on the estimation of phylogenetic parameters of *lmp7*.
- Highlight the use and pitfalls of simple and fast genotyping methods of MHC type for behaviour studies.

The approach I have taken is to isolate and sequence MHC class Ia alleles from *X. laevis*, and characterize the relatively small-scale inter-allelic recombination events within the locus. This sample will also be able to be used to estimate variable substitution rates among sites in a sequence and infer the influence, if any, of natural selection using substitution rate ratios. Where applicable, newer but currently available ML methods are used to infer molecular evolutionary parameters, and results are compared with earlier findings obtained using older pairwise methods. The mode of MHC evolution in *X. laevis* (in terms of recombination rates, substitution ratios, levels of polymorphism) is compared with humans and salmonid fishes in order to further elucidate correlations of genomic features of the MHC and differences in modes of evolution.

In addition to investigating the use of ML methods and models of evolution at the population level, I also investigate their use among phylogenetic data. For these purposes I use *lmp7* and *psmb5* to investigate multigene family evolution. A variety of vertebrate taxa and an invertebrate taxon was sampled using sequences downloaded from the Genbank database. Sampling was performed to "bracket" or "straddle" the putative duplication event by using species that diverged before the gene duplication and others that have emerged since. ML methods employing statistical models are used to infer substitution rates in time and along different

branches in a tree since the duplication event from which *lmp7* and *psmb5* were generated from their immediate common ancestral locus. These data are used to test the constancy of evolution and to detect changes in the ratios of nonsynonymous and synonymous substitution rates since duplication. Also, use of models is applied to a sample of *lmp7* homologues to investigate the effects of model complexity and selection on ML phylogenetic reconstruction.

REFERENCES

Abbas AK, Lichtman AH, Pober JS (2000) Cellular and Molecular Immunology. W. B. Saunders, Philadelphia

Akiyama K-y, Yokota K-y, Kagawa S, Shimbara N, Tamura T, Akioka H, Nothwang HG, Noda C, Tanaka K, Ichihara A (1994) cDNA cloning and interferon-γ down-regulation of proteasomal subunits X and Y. Science 265:1231-1234

Anderson DR, Burnham KP, Gould WR, Cherry S (2001) Concerns about finding effects that are actually spurious. Wildlife Society Bulletin 29:311-316

Aoyagi K, Dijkstra JM, Xia C, Denda I, Ototake M, Hashimoto K, Nakanishi T (2002) Classical MHC class I genes composed of highly divergent sequence lineages share a single locus in rainbow trout (*Oncorhynchus mykiss*). Journal of Immunology 168:260-273

Arendt CS, Hochstrasser M (1997) Identification of the yeast 20S proteasome catalytic centers and subunit interaction required for active-site formation. Proceedings of the National Academy of Sciences, USA 94:7156-7161

Beck S, Kelly A, Radley E, Khurshid F, Alderton RP, Trowsdale J (1992) DNA sequence analysis of 66 kb of the Human MHC class II region encoding a cluster of genes for antigen processing. Journal of Molecular Biology 228:433-441

Beck S, Trowsdale J (1999) Sequence organization of the class II region of the human MHC. Immunological reviews 167:201-210

Beck S, Trowsdale J (2000) The human Major Histocompatibility Complex: lessons from the DNA sequence. Annual Review of Genomics and Human Genetics 1:117-137

Bingulac-Popovic J, Figueroa F, Sato A, Talbot WS, Johnson SL, Gates M, Postlethwait JH, Klein J (1997) Mapping of MHC class I and class II regions to differnet linkage groups in the zebrafish, *Danio rerio*. Immunogenetics 46:129-134

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC (1987a) The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. Nature 329:512-518

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC (1987b) Structure of the human class I histomcompatibility antigen, HLA-A2. Nature 327:506-512

Boes B, Hengel H, Ruppert T, Molthaup G, Koszinowski UH, Kloetzel P-M (1994) Interferon-γ stimulation modulates the proteolytic activity and cleavage site

preference of 20S mouse proteasomes. Journal of Experimental Medicine 179:901-909

Bontrop RE, Otting N, de Groot N, Doxiadis GGM (1999) Major histocompatibility complex class II polymorphisms in primates. Immunological reviews 167:339-350

Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 364:33-39

Brown MG, Driscoll J, Monaco JJ (1991) Structural and serological similarity of MHC-linked LMP and proteasome (multicatalytic proteinase) complexes. Nature 353:355-357

Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference: a practical information-theoretic approach. Springer-Verlag, New York

Clark MS, Pontarotti P, Gilles A, Kelly A, Elgar G (2000) Identification and characterization of a B proteasome subunit cluster in the Japanese Pufferfish (*Fugu rubripes*). Journal of Immunology 165:4446-4452

Coux O (1996) Structure and functions of the 20S and 26S proteasomes. Annual Review of Biochemistry 65:801-847

Driscoll J, Brown MG, Finley D, Monaco JJ (1993) MHC-linked *LMP* gene products specifically alter peptidase activities of the proteasome. Nature 365:262-264

Du Pasquier L, Miggiano VC, Kobel HR, Fischberg M (1977) The genetic control of histocompatibility reaction in natural and laboratory-made polyploid individuals of the clawed toad *Xenopus*. Immunogenetics 5:129-141

Edwards SV, Wakeland EK, Potts WK (1995) Contrasting histories of avian and mammalian MHC genes revealed by class II B sequences from songbirds. Proceedings of the National Academy of Sciences, USA 92:12200-12204

Figueroa F, Gunther E, Klein J (1988) MHC polymorphism pre-dating speciation. Nature 335:265-267

Figueroa F, Mayer WE, Sato A, Zaleska-Rutczynska Z, Hess B, Tichy H, Klein J (2001) Mhc class I genes of swordtail fishes: *Xiphophorus*: variation in the number of loci and existence of ancient gene families. Immunogenetics 53:695-708

Flajnik MF, Canel C, Kramer J, Kasahara M (1991) Evolution of the major histocompatibility complex: molecular cloning of major histocompatibility

complex class I from the amphibian *Xenopus*. Proceedings of the National Academy of Sciences, USA 88:537-541

Flajnik MF, Kasahara M (2001) Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. Immunity 15:351-362

Flajnik MF, Kasahara M, Shum BP, Salter-Cid L, Taylor E, Du Pasquier L (1993) A novel type of class I gene organization in vertebrates: a large family of non-MHC-linked class I genes is expressed at the RNA level in the amphibian *Xenopus*. EMBO 12:4385-4396

Flajnik MF, Kaufman J, Riegert P, Du Pasquier L (1984) Identification of class I major histocompatibility complex encoded molecules in the Amphibian *Xenopus*. Immunogenetics 20:134-143

Flajnik MF, Ohta Y, Greenberg AS, Salter-Cid L, Carrizosa A, Du Pasquier L, Kasahara M (1999) Two ancient allelic lineages at the single classical class I locus in the *Xenopus* MHC. Journal of Immunology 163:3826-3833

Gaczynska M, Rock KL, Goldberg AL (1993) γ-interferon and expression of MHC genes regulate peptide hydrolysis by proteasomes. Nature 365:264-267

Gaczynska M, Rock KL, Spies T, Goldberg AL (1994) Peptidase activities of proteasomes are differentially regulated by the major histocompatibility complex-encoded genes for LMP2 and LMP7. Proceedings of the National Academy of Sciences USA 91:9213-9217

Glynne R, Powis SJ, Beck S, Kelly A, Kerr L-A, Trowsdale J (1991) A proteasome-related gene between the two ABC transporter loci in the class II region of the MHC. Nature 353:357-360

Goldberg AL, Rock KL (1992) Proteolysis, proteasomes and antigen presentation. Nature 357:375-379

Graser R, Vincek V, Takami K, Klein J (1998) Analysis of zebrafish *Mhc* using BAC clones. Immunogenetics 47:318-325

Gray CW, Slaughter CA, DeMartino GN (1994) PA28 activator protein forms regulatory caps on proteasome stacked rings. Journal of Molecular Biology 236:7-15

Groetterup M, Soza A, Eggers M, Kuehn L, Dick TP, Schild H, Rammensee H-G, Koszinowski UH, Kloetzel P-M (1996) A role for the proteasome regulator PA28a in antigen presentation. Nature 381:166-168

Groll M, Ditzel L, Lowe J, Stock D, Bochtler M, Bartunik HD, Huber R (1997) Structure of the 20S proteasome from yeast at 2.4A resolution. Nature 386:463-471

Gu X, Nei M (1999) Locus specificity of polymorphic alleles and evolution by a birth-and-death process in mammalian MHC genes. Molecular Biology and Evolution 16:147-156

Gyllensten U, Sundvall M, Erlich HA (1991) Allelic diversity is generated by intraexon sequence exchange at the *DRB1* locus of primates. Proceedings of the National Academy of Sciences, USA 88:3686-3690

Hashimoto K, Okamura K, Yamaguchi H, Ototake M, Nakanishi T, Kurosawa Y (1999) Conservation and diversification of MHC class I and its related molecules in vertebrates. Immunological reviews 167:81-100

Heinemeyer W, Fischer M, Krimmer T, Stachon U, Wolf DH (1997) The active sites of the eukaryotic 20 S proteasome and their involvement in subunit precursor processing. Journal of Biological Chemistry 272:25200-25209

Hess CM, Edwards SV (2002) The evolution of the major histocompatibility complex in birds. BioScience 52:423-431

Hillis DM, Moritz C, Mable BK (1996) Molecular Systematics. Sinauer, Sunderland, p 655

Hughes AL (1997) Evolution of the proteasome components. Immunogenetics 46:82-92

Hughes AL (1998) Phylogenetic tests of the hypothesis of block duplication of homologous genes on human chromosome 6, 9 and 1. Molecular Biology and Evolution 15:584-870

Hughes AL (2000) Gene duplication and MHC origins. Immunogenetics 51:982-983

Hughes AL, Hughes MK, Watkins DI (1993) Contrasting roles of interallelic recombination at the HLA-A and HLA-B loci. Genetics 133:669-680

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170

Hughes AL, Nei M (1989a) Evolution of the major histocompatibility complex: independant origin of nonclassical class I genes in different groups of mammals. Molecular Biology and Evolution 6:559-579

Hughes AL, Nei M (1989b) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proceedings of the National Academy of Sciences, USA 86:958-962

Hughes AL, Nei M (1990) Evolutionary relationships of class II major-histocompatibility-complex genes in mammals. Molecular Biology and Evolution 7:491-514

Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex loci of vertebrates. Annual Review of Genetics 32:415-435

Jakobsen IB, Wilson SR, Easteal S (1998) Patterns of reticulate evolution for the classical class I and II HLA loci. Immunogenetics 48:312-323

Jones EY, Tormo J, Reid SW, Stuart DI (1998) Recognition surfaces of MHC class I. Immunological reviews 163:121-128

Kandil E, Namikawa C, Nonaka M, Greenberg AS, Flajnik MF, Ishibashi T, Kasahara M (1996) Isolation of low molecular mass polypeptide complimentary DNA clones from primitive vertebrates. Journal of Immunology 156:4225-4253

Kasahara M (1999) The chromosomal duplication model of the major histocompatibilit complex. Immunological reviews 167:17-32

Kasahara M, Hayashi M, Tanaka K, Inoko H, Sugaya K, Ikemura T, Ishibashi T (1996) Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. Proceedings of the National Academy of Sciences, USA 93:9096-9101

Kaufman J (1999) Co-evolving genes in MHC haplotypes: the "rule" for nonmammalian vertebrates? Immunogenetics 50:228-236

Kaufman J, Milne S, Gobel TWF, Walker BA, Jacob JP, Auffray C, Zoorob R, Beck S (1999) The chicken B locus is a minimal essential major histocompatibility complex. Nature 401:923-925

Kisselev AF, Akopian TN, Woo KM, Goldberg AL (1999) The sizes of peptides generated from protein by mammalian 26 and 20S proteasomes. Journal of Biological Chemistry 274:3363-3371

Klein J, O'hUigin C, Figueroa F, Mayer WE, Klein D (1993) Different modes of *Mhc* evolution in primates. Molecular Biology and Evolution 10:48-59

Kobari F, Sato K, Shum BP, Tochinai S, Katagiri M, Ishibashi T, Du Pasquier L, Flajnik MF, Kasahara M (1995) Exon-intron organization of *Xenopus* MHC class II B chain genes. Immunogenetics 42:376-385

Lawlor DA, Ward FE, Ennis PD, Jackson AP, Parham P (1988) *HLA-A* and *HLA-B* polymorphism predate the divergence of humans and chimpanzees. Nature 335:268-271

Li W-H (1997) Molecular Evolution. Sinauer, Sunderland, MA

Lio P, Goldman N (1998) Models of molecular evolution and phylogeny. Genome Research 8:1233-1244

Lowe J, Stock D, Jap B, Zwickl P, Baumeister W, Huber R (1995) Crystal structure of the 20S proteasome from the Archeon *T. acidophilum* at the 3.4 A resolution. Science 268:355-539

McConnell TJ, Talbot WS, McIndoe RA, Wakeland EK (1988) The origin of MHC class II gene polymorphism within the genus *Mus*. Nature 332:651-654

Michalova V, Murray BW, Sultmann H, Klein J (2000) A contig map of the *Mhc* class I genomic regin in the zebrafish reveals ancient synteny. Journal of Immunology 164:5296-5305

Miller KM, Kaukinen KH, Schulze AD (2002) Expansion and contraction of major histocompatibility complex genes: a telostean example. Immunogenetics 53:941-963

Monaco JJ, Nandi D (1995) The genetics of proteosomes and antigen processing. Annual Review of Genetics 29:729-754

Namikawa C, Salter-Cid L, Flajnik MF, Kato Y, Nonaka M, Sasaki M (1995) Isolation of Xenopus *LMP-7* homologues: striking allelic diversity and linkage to MHC. Journal of Immunology 155:1964-1971

Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proceedings of the National Academy of Sciences, USA 94:7799-7806

Nonaka M, Namikawa C, Kato Y, Sasaki M, Salter-Cid L, Flajnik MF (1997) Major histocompatibility complex gene mapping in the amphibian *Xenopus* implies a primordial organization. Proceedings of the National Academy of Sciences, USA 94:5789-5791

Nonaka M, Yamada-Namikawa C, Flajnik MF, Du Pasquier L (2000) Trans-species polymorphism of the major histocompatibility complex-encoded proteasome subunit LMP7 in an amphibian genus, *Xenopus*. Immunogenetics 51:186-192

Ohta Y, McKinney EC, Criscitiello MF, Flajnik MF (2002) Proteosome, Transporter associated with antigen processing, and class I genes in the Nurse Shark *Ginglymostoma cirratum:* evidence for a stable class I region and MHC haplotype lineages. Journal of Immunology 168:771-781

Ohta Y, Okamura K, McKinney EC, Bartl S, Hashimoto K, Flajnik MF (2000) Primitive synteny of vertebrate histocompatibility complex class I and class II genes. Proceedings of the National Academy of Sciences, USA 97:4712-4717

Ohta Y, Powis SJ, Coadwell WJ, Haliniewski DE, Liu Y, Li H, Flajnik MF (1999) Identification and mapping of *Xenopus* TAP2 genes. Immunogenetics 49:171-182

O'hUigin C, Satta Y, Hausmann A, Dawkins RL, Klein J (2000) The implications of intergenic polymorphism for Major Histocompatibility Complex evolution. Genetics 156:867-877

Parham P, Ohta T (1996) Population biology of antigen presentation by MHC class I molecules. Science 272:67-74

Rock KL, Gramm C, Rothstein L, Clark K, Stein R, Dick L, Hwang D, Goldberg AL (1994) Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. Cell 78:761-771

Rock KL, York IA, Saric T, Goldberg AL (2002) Protein degradation and the generaton of MHC class I-presented molecules. In: Dixon FJ (ed) Advances in Immunology. Academic Press, San Diego, CA, p 1-70

Sato A, Figueroa F, Murray BW, Malaga-Trillo E, Zaleska-Rutczynska Z, Sultmann H, Toyosawa S, Wedekind C, Klein J (2000) Nonlinkage of major histocompatibility complex class I and class II loci in bony fishes. Immunogenetics 51:108-116

Sato K, Flajnik MF, Du Pasquier L, Katagiri M, Kasahara M (1993) Evolution of the MHC: isolation of class II B-chain cDNA clones from the amphibian *Xenopus laevis*. Journal of Immunology 50:2831-2843

Satta Y (1997) Effects of intra-locus recombination on HLA polymorphism. Hereditas 127:105-112

Seemuller E, Lupas A, Stock D, Lowe J, Huber R, Baumeister W (1995) Proteasome from *Thermoplasm acidophilum*: A threonine protease. Science 268:579-582

Shiina T, Shimizu K, Oka A, Teraoka Y, Imanishi T, Gojobori T, Hanzawa K, Watanabe K, Inoko H (1999a) Gene organization of the quail major histocompatibility complex (MhcCoja) class I gene region. Immunogenetics 49:384-394

Shiina T, Tamiya G, Oka A, Takishima N, Yamagata T, Kikkawa E, Iwata K, Tomizawa M, Okuaki N, Kuwano Y, Watanabe K, Fukuzumi Y, Itakura S, Sugawara C, Ono A, Yamazaki M, Tashiro H, Ando A, Ikemure T, Soeda E, Kimura M, Bahram S, Inoko H (1999b) Molecular dynamics of MHC genesis unraveled by sequence

analysis of the 1,796,983-bp HLA class I region. Proceedings of the National Academy of Sciences, USA 96:13282-13287

Shum BP, Guethlein LA, Flodin LR, Adkinson MA, Hedrick RP, Nehring RB, Stet RJM, Secombes C, Parham P (2001) Modes of Salmon MHC class I and II evolution differ from the primate paradigm. Journal of Immunology 166:3297-3308

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular Systematics. Sinauer, Sunderland, MA

Takahata N, Nei M (1990) Allelic genealogy under overdominant and frequency dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124:967-978

Takahata N, Satta Y (1998) Footprints of intragenic recombination at HLA loci. Immunogenetics 47:430-441

Takahata N, Satta Y, Klein J (1992) Polymorphism and balancing selection at Major Histocompatibility Complex loci. Genetics 130:925-938

Takami K, Zaleska-Rutczynska Z, Figueroa F, Klein J (1997) Linkages of *LMP, TAP,* and *RING3* with *Mhc* class I rather than class II genes in the Zebrafish. Journal of Immunology 159:6052-6060

Tanaka K, Kasahara M (1998) The MHC class I ligand-generating system: roles of immunoproteosomes and the interferon–γ–inducible proteosome activator PA28. Immunological reviews 163:161-176

Tanaka K, Tanahashi N, Shimbara N (2000) Proteasomes and MHC class I-peptide generation. In: Kasahara M (ed) Major Histocompatibility Complex Evolution, Structure, and Function. Springer-Verlag, Tokyo, p 203-212

The MHC Sequencing Consortium (1999) Complete sequence and gene map of a human major histocompatibility complex. Nature 401:921-923

Thornton JW, DeSalle R (2000) Gene family evolution and homology: genomics meets phylogenetics. Annual Review of Genomics and Human Genetics 1:41-73

Trowsdale J (1995) Both man and bird and beast: comparative organization of MHC genes. Immunogenetics 41:1-17

Vogel TU, Evans DT, Urvater JA, O'Connor DH, Hughes AL, Watkins DI (1999) Major histocompatibility complex class I genes in primates: co-evolution with pathogens. Immunological reviews 167:327-337

Wittzell H, Bernot A, Auffray C, Zoorob R (1999) Concerted evolution of two MHC class II B loci in pheasants and domestic chickens. Molecular Biology and Evolution 16:479-490

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution 19:908-917

Yeager M, Hughes AL (1999) Evolution of the mammalian MHC: natural selection, recombination, and convergent evolution. Immunological reviews 167:45-58

CHAPTER I

USING MODELS OF NUCLEOTIDE EVOLUTION TO BUILD PHYLOGENETIC TREES

ABSTRACT

Molecular phylogenetics and its applications are popular and useful tools for making comparative investigations in genetics; however, estimating phylogenetic trees is not always straightforward. Some phylogenetic estimators use an explicit model of nucleotide evolution to estimate evolutionary parameters such as branch lengths and tree topology. There are many models to choose from, and use of the optimal model for a particular data set is important to avoid a loss of power and accuracy in phylogenetic estimations. Here I review some molecular evolutionary forces and the parameters included in some common models of evolution used to interpret resulting patterns of molecular variation. I present some statistical methods of selecting a particular model of nucleotide evolution, and provide an empirical example of model selection. Statistical model selection strikes a balance between the bias introduced by some models and the increased variance of parameter estimates that results from using other models.

INTRODUCTION

The use of molecular phylogenetics has become widespread in immunological research because phylogenetic trees are an intuitive way to infer relationships among copies of a gene or among loci of a multigene family. Historically, the primary interest in constructing trees was the pattern of evolutionary relationships itself, or simply the topology of the tree. More recently however, phylogenetic trees are being generated to derive information regarding the processes responsible for the observed pattern of evolutionary relationships, and the tree topology becomes the framework upon which further inference can be drawn. As such, phylogenetics facilitates analysis of gene duplications, rates of evolution, polymorphisms, recombination, divergence of lineages and population demographics (Holder and Lewis 2003; Page and Holmes 1998). Accurate estimates of evolutionary parameters often hinge on the validity of a single phylogenetic reconstruction upon which inference is based. Inaccurate estimation of trees may lead to biased results and erroneous inference of processes or mechanism of evolution.

Several methods of estimating phylogenetic trees are available. Some of the more commonly used methods include neighbor joining (NJ)(Saitou and Nei 1987), maximum parsimony (MP) (Fitch 1970) and maximum likelihood (ML) (Felsenstein 1981). More recently, new methods that employ a Bayesian statistical approach (Larget and Simon 1999; Ronquist and Huelsenbeck 2003) have been successfully implemented, and these methods have quickly generated much interest (Holder and Lewis 2003; Huelsenbeck et al. 2001). While several differences exist, one common feature that unites NJ, ML and Bayesian methods is the use of explicit statistical models of nucleotide evolution.

In the context of phylogenetics, a model provides a framework through which the phylogenetic construction method estimates parameters used to find the preferred tree. The model represents the footprint of evolutionary phenomena that has generated the observed sequence data, such as mutation, selection, and genetic drift. The particular model selected for a data set depends on features of the data such as the level of variation and nucleotide frequencies. While it is not our intent to engage in a full review of phylogenetic methods (for reviews see Brower et al. 1996; Huelsenbeck and Crandall 1997; Nei 1996), ML, NJ and Bayesian methods generally benefit from their use of

models of evolution in terms of flexibility and performance (Swofford et al. 1996; Swofford et al. 2001).

At the outset, the reconstruction of molecular phylogenetic relationships seems a relatively simple exercise. However, the intricacies of DNA sequence evolution and the culmination of molecular forces acting on sequences can make phylogenetic inference a complex matter. The purpose of this chapter is to highlight the uses and advantages of nucleotide models in light of the complexities of evolutionary genetics. First I review aspects of DNA sequence evolution such as rates of evolution and changes in those rates through time and along the sequence. I then examine parameters of some models commonly used in phylogenetics that correspond to aspects of sequence evolution and discuss model selection and use. Finally, I present an empirical example of model selection in comparative immunology and use it to demonstrate how results can vary depending on the model being used and argue that appropriate model selection and use is critical to accurate scientific exploration of genetic information.

SEQUENCE EVOLUTION AND PHYLOGENETICS

SUBSTITUTIONS

As more DNA sequences become available, it is apparent that patterns of nucleotide changes used to construct trees are very complex. These complexities arise because of a number of factors contributing to and acting on the primary unit of sequence differences--substitutions. Substitutions can be classified as transitions (ti) or transversions (tv) (Figure 1.1). Transitions are substitutions between structurally similar nucleotides (e.g. A↔G, which are both purines), and transversions occur between dissimilar nucleotides (e.g. A↔T; purine to pyrimidine). Transitions are often observed at more than two times the rate of transversions (ti : tv > 2) even though there are twice as many

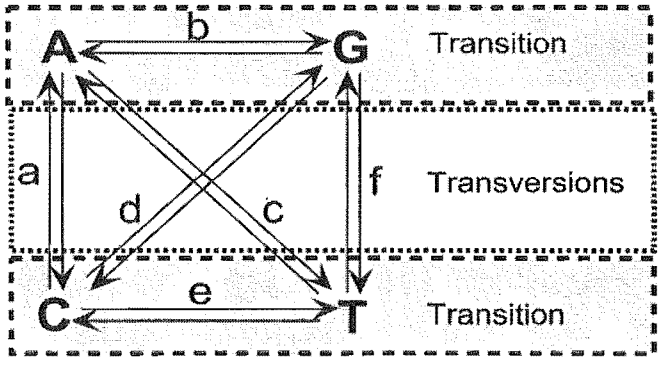


Figure 1.1. A substitution matrix representing the possible different rates of evolution for the two possible transitions, and four possible transversions (a-f). In this substitution matrix, substitution parameters are reversible, so that the rate of change from nucleotide *i* to *j* is the same as *j* to *i*.

possible transversions for any given nucleotide site. This trend towards more transitions occurs because mutation to a similar nucleotide is more likely to be tolerated than a dissimilar one, and this transition bias can be quite pronounced in some molecules, especially mitochondrial DNA. Frequently, whether or not a substitution is a transversion has implications for altering the protein coded by a DNA sequence.

Substitution rates can vary along a DNA sequence in at least two different ways. First, because of the redundant nature of the genetic code, substitutions are similarly tolerated more or less in various positions within each codon (Li 1997). For instance, the third position of a codon evolves much more quickly than the second position because substitutions at the second position usually change the amino acid encoded by that codon, while similar substitutions at the third position do not. Second, to preserve the function of the protein, its structure must be conserved in important regions; other segments of the protein may be less conserved (for a well known example, see Hughes and Nei 1988). Thus substitution rates vary in different parts of the DNA sequence correlating to different domains in the protein (i.e. among codons rather than within a codon) and can cause different parts of a gene to support different trees. The variation in substitution rates among different nucleotides in a sequence (rather than in a codon) is referred to as substitution rate heterogeneity or among-site rate variation. In a DNA sequence with among-site rate variation, some nucleotide sites undergo frequent substitutions, while others may change very slowly or not at all (Yang 1996). The occurrence of among-site rate variation alters the probabilities of nucleotide substitutions from the often-assumed notion that substitutions are randomly spread along the sequence, and is nearly ubiquitous among DNA sequences (Gu and Zhang 1997; Zhang and Gu 1998).

THE MOLECULAR CLOCK

The idea of the molecular clock is based on early observations that the number of amino acid replacements between species or lineages is proportional to the divergence time between them (Zuckerkandl and Pauling 1965). The empirical observation of a molecular clock was explained by the neutral theory of molecular evolution (Kimura 1968), where such a clock would be expected if most amino acid substitutions were selectively neutral, driven by mutations and random drift. Although the neutral theory has become pervasive in evolutionary genetics, the molecular clock does not always tick regularly (Bromham and Penny 2003). Variation of substitution rates both within a lineage and among lineages

makes the existence of a global molecular clock unlikely even though neutral mutations may dominate molecular evolution. Anything that changes the balance between drift and selection can alter the tick-rate of the molecular clock by causing a temporary increase or decrease in the number of substitutions per unit of time, and even neutral evolution can occur in an episodic manner (Ayala 1999; Gillespie 1991). Events such as gene or genome duplications, speciation or changes in the population size can change the dynamic between drift and natural selection, altering the rate of evolution if only for a short period of time.

Many lines of evidence are against a universal molecular clock; however, neutral theory still plays a prominent role in evolutionary genetics. The action of natural selection does not imply that neutral substitutions do not exist, only that they do not always accumulate with clock-like regularity. Violations of the molecular clock are commonly found in highly divergent gene sequences, genes that are the product of gene duplications (e.g. Nei et al. 1997), or genes that have experienced natural selection or changes in structure or function (e.g. Merritt and Quattro 2001). There are many difficulties associated with using a molecular clock (Arbogast et al. 2002), nevertheless, it is often the case that tests of the molecular clock (Sorhannus and Van Bell 1999) cannot reject clock-like evolution for closely related gene sequences. This could indicate that molecular evolution is clock-like for periods of evolutionary time, or that methods may lack statistical power to reject a molecular clock in some cases. Even when clock-like evolution is plausible, precise estimation of dates can still be difficult to obtain because of different assumptions and sources of uncertainty (Graur and Martin 2004). Also, methods are available that relax the assumption of a strict molecular clock and allow one to estimate evolutionary dates in lineages that have different rates (Huelsenbeck et al. 2000; Sanderson 1997; Yoder and Yang 2000).

Many evolutionary processes create irregular patterns of nucleotide substitution and the detection and characterization of these irregularities has led to a better understanding of DNA sequence evolution. In turn, our understanding of molecular evolutionary patterns has allowed us to develop statistical models used to represent the irregularities of DNA sequence evolution. For instance, through the use of these models, researchers are able to overcome common phylogenetic scenarios that are positively misleading for methods that do not use statistical models such as MP (Felsenstein 1978; Huelsenbeck and Hillis 1993; Kuhner and Felsenstein 1994). Although models are

ultimately major simplifications, summarizing many evolutionary forces and events, appropriately incorporating these models generally leads to improvement of genetic distance and phylogenetic analysis (Nei 1996).

MODELS OF NUCLEOTIDE SUBSTITUTION

PHYLOGENETIC ESTIMATORS

Neighbor Joining, ML and Bayesian methods all rely on explicit statistical models of evolution to reconstruct evolutionary trees. The NJ algorithm is different from ML and Bayesian methods because it uses the model to calculate pairwise genetic distances between sequences, and reconstructs a topology based on those distances. Maximum likelihood and Bayesian methods use the sequence data directly to reconstruct a tree, thereby utilizing information in specific nucleotide differences instead of summarizing changes with a genetic distance. Due to these differences, ML offers noteworthy statistical properties in comparison with genetic distance-based methods, but is much more computationally intensive (Huelsenbeck 1995b; Kuhner and Felsenstein 1994; Yang 1994b). While NJ and ML methods are well understood and their uses are common in the literature, Bayesian methods are relatively new.

The Bayesian method is related to ML method because they both utilize the likelihood function. However, when using Bayesian statistics to reconstruct a phylogeny, the preferred outcome is the one that maximizes the posterior probability, which is determined by the prior distribution and the likelihood of that tree. The prior distribution for trees, models and parameters can be specified to be generally uninformative to avoid bias, or it can reflect prior knowledge from other sources. Whereas other methods produce a single best estimate of evolutionary relationships and ignore uncertainty of the final outcome, Bayesian methods produce a set of trees of which the one with the highest posterior probability is accepted as the preferred tree. Bayesian methods are generally faster than ML methods, and also offer the advantage of automatically incorporating an estimate of phylogenetic uncertainty (Larget and Simon 1999). While many aspects of Bayesian phylogenetic estimation have yet to be refined and explored, these methods offer the same benefits from employing statistical models as ML and NJ (Larget and Simon 1999; Ronquist and Huelsenbeck 2003). These benefits include the flexibility to incorporate a wide range of models, easy hypothesis testing, and improvements on estimates of numbers of substitutions, efficiency and robustness (Huelsenbeck 1995a).

MODEL PARAMETERS

Statistical models of nucleotide change represent aspects of the pattern of variation that results from the process of evolution. Models vary in complexity according to the number of parameters used to represent evolutionary change. While simple models summarize nucleotide substitutions with one or two parameters, the most general models can involve more than sixty parameters (e.g. codon models that are introduced below). Model parameters can reflect differences in nucleotide frequencies, substitution rate (such as transition bias) and among-site rate variation. The substitution matrix of a model represents different rates of evolution between certain pairs of nucleotides, and the gamma distribution models among-site rate variation. In other words, the substitution matrix determines the substitution rate between specific nucleotide pairs (e.g. A ↔ G), and the gamma distribution determines the overall substitution rate at a nucleotide site. Combining different parameters has resulted in a large number of models, but many of them share several parameters (Table 1.1).

The JC69 model (Jukes and Cantor 1969) considers all possible nucleotide substitutions to have an equal probability, and is the simplest available model (Table 1.1). Felsenstein (1981) suggested a model in which probabilities of nucleotide changes were determined by the equilibrium nucleotide frequencies. Kimura (1980) proposed a model that utilizes a relatively simple substitution matrix that allows for two different rates: one for transitions and the other for transversions. Kimura (1981) and others (e.g. Tavare

Model	**Parameters**			
	Number of parameters	Nucleotide frequencies	Substitution rate in Fig. 1	Reference
JC69	1	not included	$a=b=c=d=e=f$	(Jukes and Cantor 1969)
F81	4	$\pi_A, \pi_C, \pi_G, \pi_T$	not included	(Felsenstein 1981)
K80	2	not included	$a=c=d=f, b=e$	(Kimura 1980)
K81	3	not included	$a=f, b=e, c=d$	(Kimura 1981)
HKY85	6	$\pi_A, \pi_C, \pi_G, \pi_T$	$a=c=d=f, b=e$	(Hasegawa et al. 1985)
SYM	6	not included	a, b, c, d, e, f	(Zharkikh 1994)
TrN	7	$\pi_A, \pi_C, \pi_G, \pi_T$	$a=c=d=f, b, e$	(Tamura and Nei 1993)
GTR	10	$\pi_A, \pi_C, \pi_G, \pi_T$	a, b, c, d, e, f	(Rodriguez et al. 1990)

Table 1.1. Some commonly used nucleotide models and summary of parameters. Parameters of these models can include four different base frequencies and up to six substitution rates. Flexibility of models is such that invariable sites and/or a gamma distribution can simply be added to incorporate rate variation.

1986) have also formulated models that incorporate more than two rates in the substitution matrix, thus enabling models to account for different rates of change between all of the possible nucleotide pairs. In an effort to make models more representative of empirical observations, Hasegawa et al. (1985), Felsenstein (1995), and Tamura and Nei (1993) each created models which incorporate multiple aspects of sequence evolution (Table 1.1). These models combine parameters for differences in substitution rates and differences in nucleotide frequency.

Among-site rate variation can also be incorporated into models of nucleotide evolution. The simplest way to statistically represent among-site rate variation is to divide sites into two classes: those that vary and those that are invariable. To better account for wide rate differences among the variable sites, several methods have been used (Sullivan et al. 1995; Tamura and Nei 1993), but the most successful involves the use of a gamma distribution (Gu and Zhang 1997; Yang 1993). The gamma distribution can be approximated with as little as four categories (Yang 1994a), and the statistical representation of rate variation is independent of substitution models like those described above and can simply be added to any pre-existing model (for example, we can specify a JC69+Γ model).

Under the gamma distribution, there is a continuum of probabilities of change for nucleotides, ranging from low to high. The numbers of nucleotide sites with the various rates of substitutions determines the shape of the gamma distribution that is summarized by the shape parameter (α). When most of the nucleotides are invariable or have very slow rates, then the shape of the distribution is skewed to the right (Figure 1.2). Under this scenario there are a few nucleotides with high rates and the shape parameter would be small ($\alpha < 1$), indicating a high level of rate variation, i.e. not all nucleotides evolve at a similar rate. As a result, most of the variation in the data set comes from relatively few nucleotide sites that are evolving very rapidly (substitutional "hotspots"). Large shape parameters ($\alpha > 20$) indicate a more bell-shaped distribution with most sites having intermediate rates of evolution with few nucleotides evolving at very high or low rates (Figure 1.2). As the shape parameter becomes larger, more nucleotide sites have a more similar rate of evolution and among-site rate variation becomes increasingly inconsequential (Swofford et al. 1996).

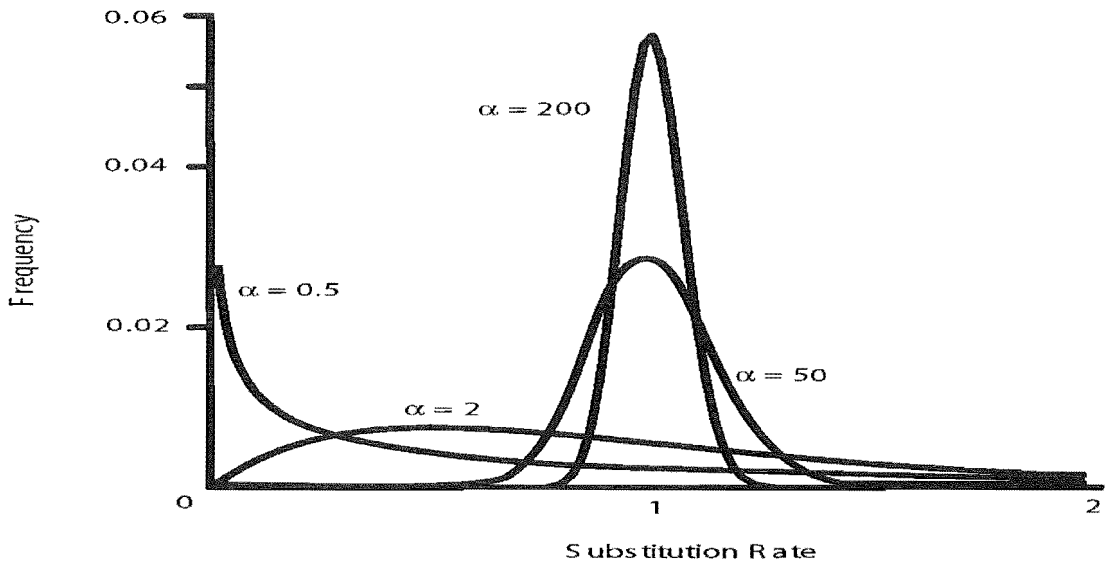


Figure 1.2. Gamma distributions calculated using different shape parameters (α). The number of nucleotides in a sequence evolving at a particular rate determines the shape parameter. When a sequence contains mostly invariable nucleotide sites and variation is concentrated at a few rapidly evolving nucleotide sites the shape parameter is small (< 1). As the proportion of variable nucleotide sites increases the shape parameter becomes larger indicating that more sites evolve at a moderate rate and fewer sites have extremely high or low rates.

The above mentioned model parameters all work at the individual nucleotide level, and therefore treat each nucleotide as an independent unit. However, for protein coding DNA sequence this is not the case. Whether or not a substitution changes an amino acid depends on the other nucleotides in that codon when the substitution occurs, thus individual nucleotide sites in protein coding sequence are not independent. To accommodate this, nucleotide models that treat a codon triplet as an independent unit have been formulated to more accurately model coding DNA (Goldman and Yang 1994; Muse and Gaut 1994; Pedersen et al. 1998). Variations of these models provide parameters to account for transition bias, codon frequency, rate variation among codon positions, and different rates for nonsynonymous substitutions (Yang et al. 2000). Codon models can become very complex by parameterizing each codon frequency, but these models can also approximate codon frequencies with fewer parameters. Unfortunately, these models are generally not implemented for use in reconstructing phylogenetic trees except when using

some Bayesian methods (Ronquist and Huelsenbeck 2003). Instead, codon models have been typically used to estimate substitution rates and detect levels of natural selection acting on a protein.

EFFECTS OF MODELS

The performance of a model-based phylogenetic method may depend on the fit of the model to the data (Huelsenbeck and Crandall 1997). Similarly, the efficiency of distance based methods is dependant on the accuracy of model-based estimates of genetic distance (Nei 1996). For sets of sequences that are long with low levels of polymorphism, the model may have little effect on the outcome of analysis. However, when working with more divergent sequences, the use of one model over another can alter the results of analysis, and even lead to strong support for the wrong tree topology (Kelsey et al. 1999), a fact that underscores the importance of using the best-fit model for a particular data set. Due to the wide diversity in size, variation and rates of evolution among different data sets, there is no single best-fit model suited for use in any data set. Use of inadequate, overly simplistic models selected without statistical validation often leads to biased estimation of evolutionary genetic parameters (Buckley et al. 2001; Gu and Li 1996; Huelsenbeck 1995a; Huelsenbeck and Hillis 1993; Swofford et al. 2001).

The model parameter with one of the strongest influences on genetic distance and phylogenetic estimation is among-site rate variation. Rate variation among sites is particularly problematic and misleading when substitution rates also vary among branches in the tree (e.g. nonclock-like evolution) (Kuhner and Felsenstein 1994). When both types of variation are present, use of the best fit model seems to be essential to obtain the correct tree topology (Cunningham et al. 1998; Yang 1996). Except in cases with strong rate variation among both sites and lineages, tree topology estimation is relatively robust to violations of model assumptions (Yang 1994b; Yang et al. 1995). Unfortunately the same robustness does not extend to estimation of parameters such as substitution rates, branch lengths and genetic distance. Failing to include rate heterogeneity among sites results in underestimation of the number of substitutions at highly mutable sites (Yang 1996). Consequently, branch lengths are underestimated, and this effect is much more prominent in longer branches than shorter ones (Buckley et al. 2001). This is likely to be due to the fact that phylogenetic estimators give greater weight to highly variable sites in a sequence (Yang 1994a).

Simplifying the assumptions of a model by failing to include a factor for transition bias can also adversely alter the outcome of analysis. A transition bias is found universally among DNA sequences (Wakeley 1994) and inclusion of this parameter is essential for accurate estimates of genetic distance for NJ analysis (Tajima and Takezaki 1994; Tamura 1992). Similarly, failure to incorporate transition bias will result in underestimation of branch lengths in ML phylogeny estimation (Yang et al. 1994). Aside from the inherent problems of branch length and genetic distance underestimation, these factors can alter the tree topology and lead to erroneous conclusions regarding the dates of lineage splitting (Tamura and Nei 1993). There is also an interplay between transition bias and among-site rate variation, so that the level of among site rate variation is underestimated (overestimation of α) using models that exclude a transition bias (Yang et al. 1994).

One of the major advantages of using models is the ability to more accurately estimate the actual number of substitutions that have occurred in a set of sequences. This allows researchers to include sequences of high variability because homoplasy in the form of superimposed substitutions can be accounted for with the use of models. The alternative way of dealing with sites or sequences which are suspected of saturation of substitutions is simply to eliminate then from consideration. While this does effectively eliminate the influence of homoplasy at those sites, any information that can be gleaned from those sites is also lost and the size of the sample is decreased, exposing the analysis to the increasing effects of sampling error or bias.

While potential problems with simple models are documented, some also dispute the utility of more general models (Sanderson and Kim 2000). Some criticisms of very complex models point out that these models have greater difficulty distinguishing between tree topologies because of smaller differences in likelihood scores, and that as more model parameters are added, more error is associated with each parameter estimate. These properties of complex models are general statistical phenomena and are not limited to phylogenetic analysis; however, while these points are valid, they arise because of random rather than systematic error. As a result these problems can be mediated rather than aggravated by addition of data (Swofford et al. 1996). The amount of data required for consistent phylogenetic analysis depends on the shape of the tree, numbers of taxa and levels of diversity. If the tree shape is not symmetric and branch lengths are very long, then analysis of data with less than 500 nucleotides will generally not be reliable,

especially for more general models (Huelsenbeck and Hillis 1993; Sullivan and Swofford 2001). Consistency and reliability of phylogenetic inference is expected to increase by analyzing longer sequences and additional taxonomic sampling.

The potential bias introduced through using a particular model also has an effect upon the level of support given to a tree topology with techniques like bootstrapping (Felsenstein 1985). The most widely accepted interpretation of the bootstrap is that it is the level of support for a particular node of a tree that the data provides (Hillis and Bull 1993). As such, it represents whether the same topology might be recovered if more data are collected, rather than if the relationship is correct. However one interprets the bootstrap values, the accuracy and precision of the bootstrap values depends on the fit of model (Buckely and Cunningham 2002). For instance, if a phylogenetic method or a model is used that has systematic bias, then the bootstrap will also reflect that bias (Swofford et al. 1996). Consequently, bootstrap values used in such a case will be artificially high and reflect strong support for incorrect branching patterns.

Bayesian methods estimate a level of phylogenetic support that is seen as an intuitive measure of uncertainty regarding each tree topology. Less work has been done to evaluate Bayesian measures of support and the relationship of model specificities and levels of support (Lemmon and Moriarity 2004). However, some research shows that Bayesian measures of support are good estimates of phylogenetic accuracy (Wilcox et al. 2002), but others conclude that these values are overestimates of the true level of uncertainty (Simmons et al. 2004; Suzuki et al. 2002). Regardless of the procedure used to measure phylogenetic support, caution interpreting results is warranted and use of a statistically rigorous method of selecting a model is recommended.

Although conflicting examples of model complexity and phylogenetic accuracy can be found (Buckely and Cunningham 2002; Takahashi and Nei 2000), one trend that has emerged is that because of the increase in variance, very short sequences (which are statistically equated with small sample sizes) often do not support the use of the same level of model complexity as longer sequences. Even though the underlying evolution of short sequences may be just as complex as longer sequences, the larger variance inherent with generalized models and small sample sizes makes these types of data more prone to the effects of over-parameterization (Burnham and Anderson 2002). While the relationship of model parameters and performance of Bayesian, ML and NJ tree estimation is not always straightforward, a trade-off between the bias of simple models and the increased variance

of more general models is generally observed (Swofford et al. 2001). Consideration of models should take into account the size of the data set, level of divergence, amount of differences in substitutions between different nucleotides, and constancy of rate of evolution both in time and along the sequences. Use of objective criteria to select models will help avoid problems associated with model over-fitting by ensuring that models are not excessively complex and avoid phylogenetic bias by selecting more realistic models (Posada and Crandall 2001a).

Models that can be implemented in popular phylogenetics programs such as PAUP* (Swofford 1998), PHYLIP (Felsenstein 1995), MEGA2 (Kumar et al. 2001) and MRBAYES (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) are useful approximations of DNA sequence evolution. Use of one particular model versus another often changes the outcome of analysis, and the choice of models can be more important than the method of phylogenetic reconstruction. Given that the model plays a large role in the results of analysis, it seems that the choice of one model over another should be justified in some way. Unfortunately, it is still commonplace for models to be used indiscriminately and without justification. The question then becomes, which model is appropriate for a particular data set and how can that model be justified?

MODEL SELECTION AND USE

To minimize adverse effects of model over-fitting and model under-fitting, the ideal use of models is to incorporate as much model complexity as needed and no more. Fortunately, methods for selecting the most appropriate model for a particular data set have been proposed. These methods provide a rigorous statistical framework in which to select and justify the best fit model. With the goal of finding the simplest model that accurately approximates sequence evolution, Rzhetsky and Nei (1995) developed statistics for selecting models. These tests are independent of evolutionary time and do not require an *a priori* phylogeny on which to base inference. While this method is computationally efficient, its application is model-specific and restricted to a limited subset of the available models.

Another method is to use the Likelihood Ratio Test (LRT) to compare models (Huelsenbeck and Crandall 1997). The LRT statistic is calculated by obtaining the likelihood scores of a null model (L_0) and an alternative model (L_1). The two scores are then compared by taking twice the difference in the logarithm of the likelihoods to obtain

the statistic [$\delta = 2(\ln L_1 - \ln L_0)$]. Use of the LRT in phylogenetics is commonplace for hypothesis testing and the distributions and performance of the test have been investigated (Goldman et al. 2000; Whelan and Goldman 1999). When the models compared are nested (one is a special case of the other), the Chi-square distribution (χ^2) is a good approximation of the null distribution of the LRT statistic (df = the difference in the number of free parameters in the two models). In some special cases, fixing one of the parameters of the more parameter-rich model at either boundary (0 or ∞) reduces the model to the simpler null model, and a mixed distribution is used (Goldman and Whelan 2000).

The LRT can be performed on any of the available models, but it requires an *a priori* input phylogeny to estimate the likelihood of the models (Posada and Crandall 2001a). It is also easy to test several models against each other in a series of LRTs that can be performed in a hierarchical fashion (Figure 1.3). The likelihood scores of the two models are compared using the LRT test statistic, δ, and significance of the LRT statistic is determined. The better fitting model is retained, it becomes the null model, and the process is iterated with successively more general models of evolution until the addition of further complexity in the alternative model does not create a significantly better fit to the data (Figure 1.3 and Table 1.2). The LRT may be appealing, but the significance of LRTs are easily calculated only for nested models and the *a priori* distribution of significance for non-nested comparisons is not well established. Performance tests of the LRT also show that this criterion is good at recovering the model used to simulate the sequence data (Posada and Crandall 2001a), although we should keep in mind that in reality the true model of nucleotide substitution is unknown, and it is much more complex than any candidate model that we can select.

null model	alternative model	parameter tested	LRT (δ)	P value
JC69	F81	equal base frequencies	4.680	0.196
JC69	K80	ti = tv	90.022	0.000
K80	SYM	equal ti and tv rates	50.340	0.000
SYM	SYM + Γ	equal rates among sites	314.512	0.000
SYM + Γ	SYM + Γ + I	no invariable sites	14.866	0.000

Table 1.2. Several models are compared successively to determine the best fitting model for a data set, starting with the simplest model and increasing complexity. The parameter being tested is assumed by the current null model but not the alternative model. The null model is rejected when the *P*-value of the LRT is < 0.01 using a χ^2 or mixed χ^2 distribution.

Another way of selecting the most appropriate model for a data set is to use the Akaike information criterion (AIC) (Akaike 1974), which can be thought of as the amount of information lost when a particular model is used to approximate reality. The AIC implements best-fit model selection by calculating the likelihood of proposed models, and imposing a penalty based on the number of model parameters. Parameter-rich models incur a larger penalty than more simple models so that fitting an excessively complex model is not likely. The best fitting model is the one with the smallest AIC value, (AIC = $-2 \ln L_i + 2 N_i$), where L_i is the likelihood for model i and N_i is the number of free parameters in model i. Although the use of LRTs is much more extended in phylogenetics than the use of the AIC, the latter offers important advantages (Burnham and Anderson 2002). The AIC is able to compare non-nested models and simultaneously compares all candidate models, rather than performing sequential pair-wise comparisons; the AIC also has a simple adjustment that more heavily penalizes complex models for data comprised of small samples (i.e. short sequences). The AIC also allows for model selection uncertainty and model averaging. In addition, the AIC recognizes that the true model is not among the set of candidate models so it tries to find the candidate model that best "approximates" the true unknown model of molecular evolution given the amount of information in the data. The objective of model selection is to find the model that will avoid bias and excessive variance. The model that is best suited to that end will not be an exact representation of cumulative evolutionary processes, but a useful approximation that is appropriate for the level of polymorphism and size of the data set.

Bayesian statistics have also been adapted for use in phylogenetic model selection. Bayes factors make pairwise model comparisons and are therefore analogous to the LRT procedure (Huelsenbeck et al. 2004; Suchard et al. 2001). Alternatively, the Bayesian Information Criterion (BIC) can be used (Schwarz 1978). This method more easily enables comparisons of multiple models and is easy to calculate. The posterior probabilities of Bayesian statistics are already used to discriminate between phylogenetic trees and these measures can also be used to choose among multiple models (Raftery 1996). Like the AIC, Bayesian methods allow estimation of model uncertainty and allow estimation of a phylogeny using a set of candidate models in a model averaging procedure. An important distinction of Bayesian statistics is that calculation of likelihoods proceeds differently, so that likelihood values compared using Bayesian methods are different from those used in AIC or LRT comparisons.

The above techniques compare model fitness relative to other candidate models, but measuring overall adequacy of a model can also be done. To do this, Navidi et al. (1991) and Goldman (1993) describe a test that compares a model with an unconstrained model and the appropriate distribution to test significance. Also, Bayesian methods have recently been adapted to examine the adequacy of models (Bollback 2002). While the unconstrained model is very complex, it is worth noting that when comparing any two models, only aspects in which the models differ are tested. Any aspects models have in common or aspects that are not included in either model remain untested. The outcome of general adequacy tests may find that the selected model is not a complete representation of the data. This is usually thought to be the result of the stringency of the test, instead of gross misrepresentation of the data by the model. Rather, this outcome simply means that the model does not perfectly describe all of the underlying processes of molecular evolution, as would be expected.

The impact of models on phylogenetic analysis is very significant, strongly affecting branch lengths and often topology as well. The use of any particular model is not wrong *per se*, but I advocate statistical, objective selection among available candidate models to maximize the use of available models for each data set. Unfortunately the model used for analysis is often not justified or even reported in the literature despite its influence on the outcome. However, easy-to-use computer programs that implement rigorous statistical selection of models are available (Posada and Crandall 1998). In the following I demonstrate their use and show how model selection determines the outcome of phylogenetic analysis.

EMPIRICAL EXAMPLE

DATA

To illustrate aspects of model selection, I reconstructed the phylogenetic relationships of nine taxa using DNA sequences of the *lmp7* gene downloaded from the Genbank database (accession numbers: human, *Homo sapiens* BC001114 (the human *lmp7* is also termed PSMB8 or RING10); mouse, *Mus musculus* U22032; frog, *Xenopus laevis* D44540; salmon, *Salmo salar* AF184938; zebrafish, *Danio rerio* AF032390; medaka, *Oryzias latipes* D89725; pufferfish, *Fugu rubripes* AJ271723; nurse shark, *Ginglymostoma cirratum* D64057; horn shark, *Heterodontus francisci* AF363583). Copies of the gene are from a variety of vertebrates from which full length cDNA was obtained, and the

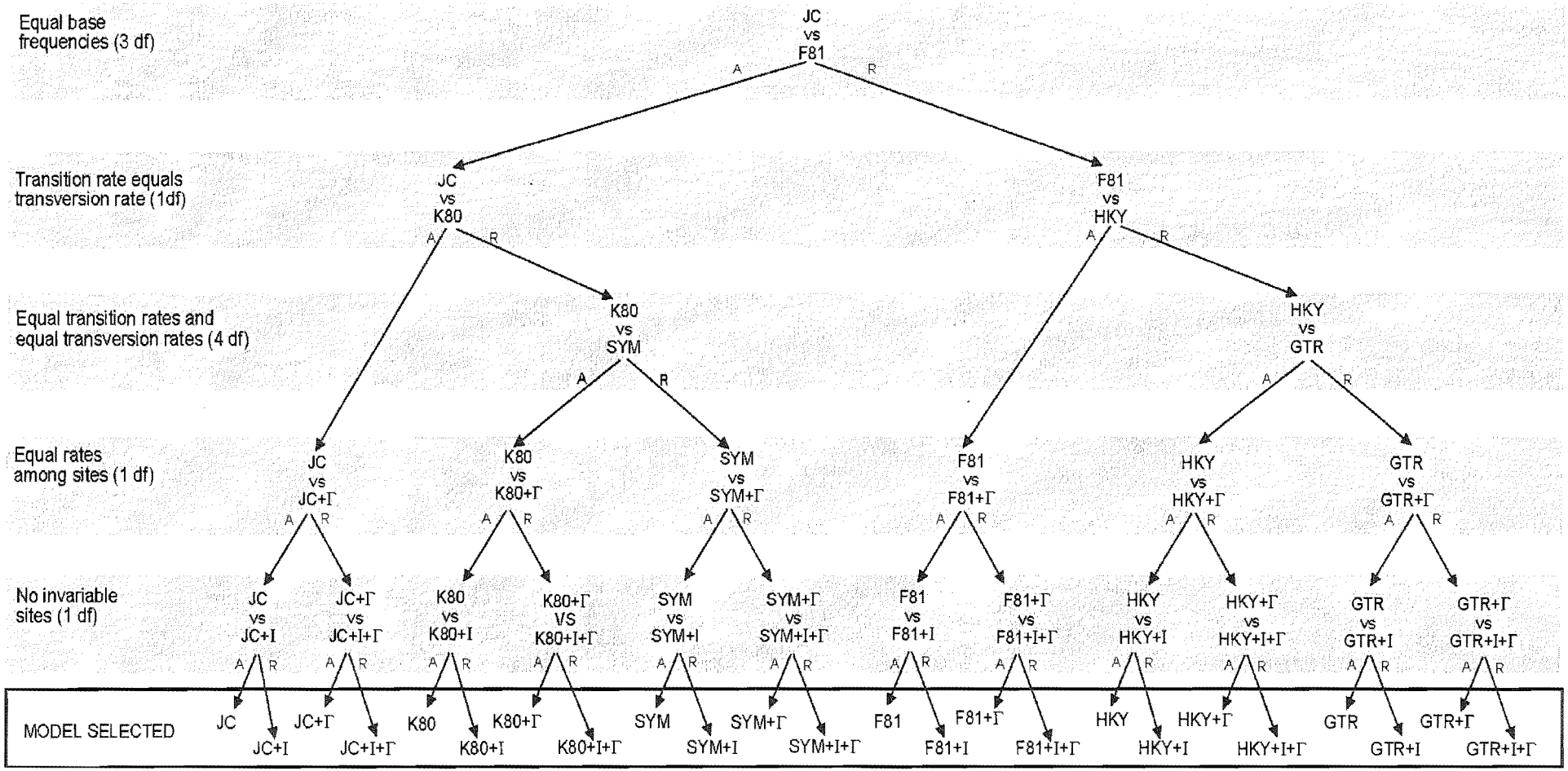


Figure 1.3. A "decision tree" of a hierarchical likelihood ratio test. Hypotheses tested are indicated on the left, and this schematic begins with the simplest model and progresses to more complex models in a stepwise manner. The pathway chosen depends on acceptance or rejection of LRT scores, based on a chi squared distribution ($P < 0.01$). In order to preserve the clarity of the figure, not all available models are shown. Models depicted here are: JC (Jukes and Cantor 1969); F81 (Felsenstein 1981); K80 (Kimura 1980); HKY (Hasegawa et al. 1985); SYM (Zharkikh 1994); GTR (Rodriguez et al. 1990). I = proportion of invariable sites; Γ = gamma distribution of rates among sites.

estimated phylogenetic relationships could be used in the framework of studying multigene family evolution, estimating substitution rates, or establishing homology of gene copies. The leader peptide was excluded from analysis, leaving only the coding sequence from the mature protein. The sequences were aligned using Clustal W (Thompson et al. 1994) and alignments were inspected to ensure that the integrity of the coding frame was preserved. The best-fit model for these data was selected using the LRT and AIC after calculating likelihood scores of 24 models using PAUP*4.0 (Swofford 1998).

METHODS

The best fitting model for these data was evaluated according to a hierarchical LRT. The AIC method of model selection was also used to find the best-fit model by calculating the likelihood and subsequently the AIC score of all models. Phylogenetic trees were calculated using 24 models of evolution selected to represent a variety of statistical complexity. These models have an arbitrary relationship to the data, and the resulting trees can be compared to those obtained using models selected using rigorous statistical criteria. Here I use the ML method of phylogenetic construction as implemented in PAUP* (Swofford 1998) because it is known to be robust to violations of model assumptions and because the statistics of ML estimation are well understood (Huelsenbeck and Crandall 1997; Yang 1994b; Yang et al. 1995). I calculated these scores manually to demonstrate the method, but the program MODELTEST (Posada and Crandall 1998) provides the appropriate command block for PAUP* to automatically calculate the likelihood scores for 56 models, which can then be automatically compared using the LRT and AIC in the MODELTEST program.

RESULTS

The model selected by both the LRT and AIC is the SYM model with both invariable sites and a gamma distribution of among-site rate variation (SYM + Γ + I; see Tables 1.2 and 1.3) (Tamura and Nei 1993; Yang 1994a). This model includes a substitution matrix allowing for 6 different rates of substitutions: one for each type of reversible nucleotide change. There is no significant heterogeneity of nucleotide frequencies accounted for in the model, but the model makes provisions for considerable rate variation along the gene sequence (see Table 1.3). The invariable sites of the sequence alignment are accounted for

Parameter	value
Substitution matrix	
A : C	1.49
A : G	2.12
A : T	1.53
C : G	0.62
C : T	4.24
G : T	1.00
base frequencies	
A	0.25
C	0.25
G	0.25
T	0.25
proportion invariable sites	0.411
Gamma shape parameter	2.827

Table 1.3. Molecular evolutionary parameter values of best-fit model, SYM + Γ + I, selected under the LRT and AIC criteria.

in the model and the gamma distribution represents rate heterogeneity only among variable sites. As the distribution of the gamma shape parameter is skewed towards the right, most of the variable nucleotides evolve fairly slowly, with a few sites evolving more rapidly. Models with few parameters commonly used to reconstruct phylogenetic relationships were rejected by both selecting criteria in favour of more general models (Table 1.2).

A test of the overall adequacy of the preferred model against the unconstrained model (Goldman 1993) indicates a sufficient level of support for the SYM + Γ + I model. The test statistic of the difference in likelihoods between the unconstrained and SYM + Γ + I models was 1091.546. Monte Carlo simulations under the null (SYM + Γ + I) model hypothesis were done to determine the null distribution of differences in likelihood between the unconstrained and SYM + Γ + I models. This distribution ranged from 946.723 to 1196.620 with a mean value of 1069.492. The test statistic falls well within the 95th percentile of the distribution, indicating that the null hypothesis (SYM + Γ + I) cannot be rejected against the unconstrained model under these criteria. The best-fit model selected above was therefore used to reconstruct the phylogenetic relationships among these taxa, and the result indicates a topology consistent with generally accepted relationships (Figure 1.4).

When the evolutionary relationships among these genes were estimated using other models, three different tree topologies emerged (models and likelihood scores found in Table 1.4). Many simple models rejected by statistical model selection criteria preferred a tree in which the frog and shark share a most recent common ancestor, and this clade is a sister group to a clade in which mammals and teleost fish share a most recent common ancestor (Figure 1.5a). Eight of the nine models that reproduce this topology share a common feature: they do not have a substitution matrix specifying different rates for substitutions between different nucleotide pairs. Seven other models reconstructed a tree

Model	-lnL	δAIC	Topology
JC	3813.934	455.742	Figure 1.5a
JC + I	3652.954	135.782	Figure 1.5a
JC + Γ	3659.403	148.680	Figure 1.5a
JC + I + Γ	3650.366	132.606	Figure 1.5a
F81	3811.594	455.062	Figure 1.5a
F81 + I	3651.362	136.598	Figure 1.5a
F81 + Γ	3657.217	148.308	Figure 1.5a
F81 + I + Γ	3648.494	132.862	Figure 1.5a
K80	3768.922	367.718	Figure 1.5b
K80 + I	3602.037	35.948	Figure 1.4
K80 + Γ	3606.883	45.640	Figure 1.5a
K80 + I + Γ	3598.376	30.626	Figure 1.5b
HKY	3766.357	368.588	Figure 1.5b
HKY + I	3601.262	40.398	Figure 1.4
HKY + Γ	3605.544	48.962	Figure 1.5b
HKY + I + Γ	3597.331	34.536	Figure 1.4
SYM	3743.752	325.378	Figure 1.4
SYM + I	3583.904	7.682	Figure 1.4
SYM + Γ	3586.496	12.866	Figure 1.5b
SYM + I + Γ	3579.063	Best	Figure 1.4
GTR	3737.487	318.848	Figure 1.5b
GTR +I	3582.327	10.528	Figure 1.4
GTR + Γ	3583.892	13.658	Figure 1.5b
GTR + I + Γ	3576.966	1.806	Figure 1.4

Table 1.4. Log Likelihood scores (-ln*L*) of models calculated using the single NJ tree topology used in the hLRT. Significance of likelihood comparison summarized in Table 1.2. Topology reconstructed under each of 24 models representing various levels of complexity. See the caption of Figure 1.3 for model references.

in which frog and mammals formed a tetrapod clade and fishes formed a monophyletic group. However in this tree, the pufferfish and medaka, generally considered derived fishes, are found at the root of the teleost clade, displacing the more primitive salmon and zebrafish (Figure 1.5b). In total, eight models preferred the "correct" topology; however no clear pattern of which models reconstruct the "correct" tree exists for these data (Table 1.4). For example, not all models that include a parameter for among-site rate variation result in the "correct" tree, and some models that are more complex than the best-fit model found the "correct" tree and some reconstructed another topology. The lack of a clear pattern in progression of model parameters and tree structure illustrates that it is often impossible to tell *a priori* which models will find the same tree as the best-fit model, a fact that underscores the importance of finding the best-fit model.

DISCUSSION

It is difficult to fully assess why some models reconstruct a topology inconsistent with generally accepted taxon relationships in this example, and multiple factors of sequence evolution are often the cause. In this case, the level of diversity may contribute to misleading results. A very high level of diversity means that many potential substitutions may be unaccounted for using simple models that consistently underestimate the number

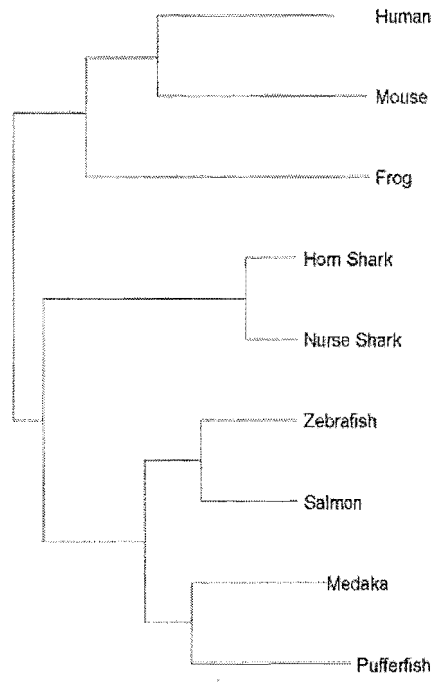


Figure 1.4. Unrooted phylogenetic tree generated using the maximum likelihood optimality criterion and the preferred model of nucleotide evolution (SYM + Γ + I) selected by the hierarchical likelihood test and the AIC criterion.

of substitutions for distantly related species (Yang et al. 1994). Multiple substitutions at given sites may provide conflicting evidence for various relationships, weakening support for a clade or overall branching pattern. This lack of consistent support renders trees with different topologies statistically indistinguishable. I tested the statistical difference among trees using the Shimodaira-Hasegawa test (1999), and found no significant difference between all three topologies (Figure 1.5a, $P = 0.305$; Figure 1.5b, $P = 0.572$). Since some more complicated models also fail to reconstruct the widely accepted "true" phylogeny it is likely that other factors play a role in misleading phylogenetic analysis. For these data, other factors such as different rates of evolution in part of the tree may also decrease the usefulness of models.

Use of simplistic models in evolutionary genetics can be misleading if the sequences do not evolve according to a molecular clock (Rzhetsky and Sitnikova 1996). I used a simple LRT to test whether or not these sequences evolve according to a molecular clock (Felsenstein 1981) to see if this may be a misleading factor for simpler models. The LRT statistic was 137 ($P < 0.0001$; df = 7) indicating that the model enforcing a strict molecular clock was a much worse fit to these data. These results corroborate those of Takezaki et al. (2002), who found variable substitution rates among lineages of proteasome components. Most statistical models of nucleotide evolution are "stationary," in that model parameters are constant across the entire tree; however, nonstationary models have been formulated that allow parameters to change with time (Gu and Li 1998;

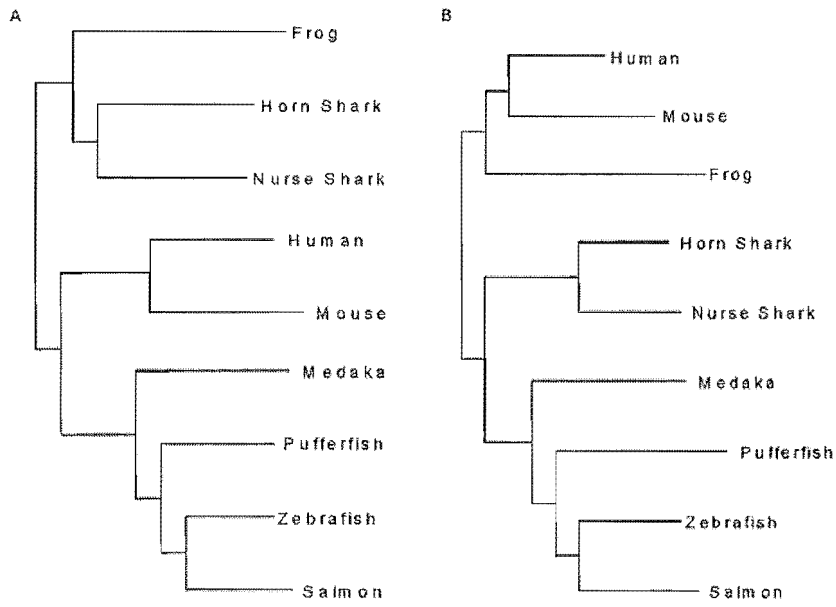


Figure 1.5. Unrooted phylogenetic trees generated using the maximum likelihood optimality criterion. Twenty-four different models of nucleotide evolution (Table 4) were systematically selected to represent a range of models with differing levels of complexity, but arbitrarily selected with regard to how well they fit the data. These models were then applied to the data, and several of thesemodels supported trees with topologies that differed from that reconstructed using the optimal model.

Huelsenbeck and Nielsen 1999). Use of these models generally improves the fit to the data and performance of the method, but greatly increased model complexity. Here, the overall rate of evolution is different among branches of the tree, therefore this and other simplifying assumptions may affect model fitness and utility.

Other factors may be involved in the failure of some models, as Whelan et al. (2001) indicate that positive or negative selection may be an unaccounted for dynamic that affects phylogenetic reconstructions. For instance, selective pressures can result in convergent evolution, causing divergent taxa to appear closely related. The test of model adequacy indicates support for the SYM + Γ + I model, but both models in that comparison make no provisions for natural selection and they both assume that data at each site is independent and identically distributed. Therefore, neither of these aspects of sequence evolution is evaluated in that comparison. Due to the coding nature of these

sequences it is likely that both natural selection and non-independence of nucleotide sites are prominent features of sequence evolution in these data.

The phylogenetics of proteasome components have been studied by others who included entire gene families in their sampling (Hughes 1997; Takezaki et al. 2002). Previous phylogenetic work on proteasome components analyzed amino acid sequences, which can be an effective means of determining phylogeny in highly divergent data. These analyses employ a variety of methods including MP, a non-model based algorithm, and NJ with a Poisson corrected distance. (The Poisson amino acid model is analogous to the JC69 nucleotide model and assumes that all changes between amino acids occur at the same rate and all amino acids are found in equal frequency.) The JTT amino acid model (Jones et al. 1992) is also used to calculate maximum likelihood scores of three preset fixed topologies. The JTT model is more suited to the analysis of divergent amino acid sequences and is based on substitution rates in a large sample of related proteins (Jones et al. 1992). In these cases, the use of a particular model is reported but no tests were conducted to select from a suite of available amino acid models (e.g. Kishino et al. 1990; Whelan and Goldman 2001). Statistical theory is often utilized through model-based phylogenetics, but the fuller potential and benefits of statistical analysis remains unemployed by not considering recent advancements in model selection. Many other examples of using evolutionary models for phylogenetic reconstruction without statistically evaluating the fit of a model are widespread in the literature (Posada and Crandall 2001a).

Another example of differing trees obtained with differing phylogenetic methodologies can be found in a study of antigen receptors by Richards and Nelson (2000). They used two methods, MP and NJ, to reconstruct the evolutionary history of members of the immunoglobulin superfamily of genes using amino acid sequences. For NJ distance calculations, they do not specify which model of evolution was used to estimate genetic distances or mention how that model was selected. However, in their analysis the model-based NJ method outperformed the MP because more monophyletic clades reflected current immune receptor classifications established by function (Richards and Nelson 2000). Even with a model of evolution, strong bootstrap support for many of the nodes in their analysis is lacking. Such a lack of support or conflicting trees may be expected when the natural limitations of protein length and ancient divergence constrain the size and signal of the sequence alignment used for analysis. Also, similar structures

and function in families of genes can cause convergence at the molecular level. Finally, the period following the gene duplications that create multigene families is often marked with increased substitution rates or varying levels of natural selection (Moore and Purugganan 2003). The temporal and often temporary change in evolutionary process makes phylogenetic analysis with stationary models of evolution more difficult.

Other data sets will have different properties that play an important role in determining the best-fit model, and population data collected from a single species presents unique obstacles for evolutionary analysis. The phylogenetic methods discussed here are designed for use on hierarchically ordered data (each sampling unit has only a single ancestor) such as the creation of two species from one. Their use on population-based sampling from a single species presents other difficulties which may further complicate analysis and create misleading results, even with correct use of statistically justified models (Posada and Crandall 2001b). For instance, in a population sample, sequences may not be related in a hierarchical manner (each unit has two ancestors (parents) in a sexually reproducing species). Further, processes at the population level, such as recombination, result in the problem that different parts of a DNA sequence have different evolutionary histories, and cannot accurately be represented by a single phylogenetic tree (Schierup and Hein 2000). Use of different parts of recombining trees typically leads to different trees that may not be correlated, depending on the relationship of the sequences that exchanged genetic information (Posada and Crandall 2002). Recombination also alters estimates of mutation rates, dating of evolutionary events, and estimates of among-site variation (Satta et al. 1999; Schierup et al. 2001). Extra effort should be taken when using phylogenetic methods to analyze data from a single species to avoid pitfalls introduced by population-level processes, and methods designed for this purpose should be employed (Posada and Crandall 2001b).

SUMMARY

The estimation of phylogenetic trees or genetic distances is a complex statistical problem in which elements such as rate of evolution, branch length, and tree topology are represented by parameters in a model (Yang et al. 1995). A phylogenetic tree and model parameters should be considered a hypothesis of evolutionary relationships statistically supported by particular data. It is important to ensure that any conclusions from evolutionary genetic analysis be as strongly supported as possible by using statistically

relevant models. Results obtained using arbitrarily selected models may easily be contradicted simply by using different models that lend support to different hypotheses (Cunningham et al. 1998; Kelsey et al. 1999). When model-based methods are used, their performance is optimized when the best model is used (Huelsenbeck 1995a), thereby lending more credibility to results obtained using statistically justified models. Some estimate of the fit of the model to the data should be calculated and used to select among available models rather than relying heavily on the robustness of the reconstruction method (Posada and Crandall 2001a). It is our position that statistical accuracy should not be sacrificed for the sake of ease or computational speed. Advancements in the statistics of model selection have already benefited every scientific discipline that uses model-based analysis. Evolutionary genetic analysis is also experiencing similar progress from these advancements. New models and improved implementation, along with selection of models under a statistically rigorous framework will continue to enhance understanding of evolutionary patterns and processes underlying the variation found in genes of the immune system.

ACKNOWLEDGEMENTS

I would like to thank Louis Du Pasquier and Ashley Sparrow for helpful comments on an earlier version of this manuscript. Janae Bos and Matt Walters provided helpful editorial and digital imagery assistance.

REFERENCES

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control AC 19:716-723

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB (2002) Estimating divergence times from molecular data on population genetic and phylogenetic time scales. Annual Review of Ecology and Systematics 33:707-740

Ayala FJ (1999) Molecular clock mirages. BioEssays 21:71-75

Bollback JP (2002) Bayesian Model Adequacy and Choice in Phylogenetics. Molecular Biology and Evolution 19:1171-1180

Bromham L, Penny D (2003) The modern molecular clock. Nature Reviews Genetics 4:216-224

Brower A, DeSalle R, Vogler AP (1996) Gene trees, species trees, and systematics: a cladisitic perspective. Annual Review of Ecology and Systematics 27:423-450

Buckely TR, Cunningham CW (2002) The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. Molecular Biology and Evolution 19:394-405

Buckley TR, Simon C, Chambers GK (2001) Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Systematic Biology 50:67-86

Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference: a practical information-theoretic approach. Springer-Verlag, New York

Cunningham CW, Zhu H, Hillis DM (1998) Best-fit maximum likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution 52:978-987

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology 27:401-410

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368-376

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791

Felsenstein J (1995) PHYLIP (Phylogenetic inference package). University of Washington, Seattle, WA

Fitch WM (1970) Toward defining the course of evolution: minimal change for a specific tree topology. Systematic Zoology 20:406-416

Gillespie JH (1991) The Causes of Molecular Evolution. Oxford University Press, New York

Goldman N (1993) Statistical tests of models of DNA substitution. Journal of Molecular Evolution 36:182-198

Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood-based tests of topologies in phylogenetics. Systematic Biology 49:652-670

Goldman N, Whelan S (2000) Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Molecular Biology and Evolution 17:975-978

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11:725-736

Graur D, Martin W (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends in Genetics 20:80-86

Gu X, Li W-H (1996) A general additive distance with time-reversibility and rate variation among nucleotide sites. Proceedings of the National Academy of Sciences, USA 93:4671-4676

Gu X, Li W-H (1998) Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proceedings of the National Academy of Sciences, USA 95:5899-5905

Gu X, Zhang J (1997) A simple method for estimating the parameter of substitution rate variation among sites. Molecular Biology and Evolution 14:1106-1113

Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape split by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22:160-174

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Systematic Biology 42:182-192

Holder M, Lewis PO (2003) Phylogeny estimation: traditional and Bayesian approaches. Nature Reviews Genetics 4:275-284

Huelsenbeck JP (1995a) Performance of phylogenetic methods in simulation. Systematic Biology 44:17-48

Huelsenbeck JP (1995b) The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of the maximum likelihood over neighbor joining. Molecular Biology and Evolution 12:843-849

Huelsenbeck JP, Crandall KA (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. Annual Review of Ecology and Systematics 28:437-466

Huelsenbeck JP, Hillis DM (1993) Success of phylogenetic methods in the four-taxon case. Systematic Biology 42:247-264

Huelsenbeck JP, Larget B, Alfaro ME (2004) Bayesian Phylogenetic Model Selection Using Reversible Jump Markov Chain Monte Carlo. Molecular Biology and Evolution 21:1123-1133

Huelsenbeck JP, Larget B, Swofford DL (2000) A compound process for relaxing the molecular clock. Genetics 154:1879-1892

Huelsenbeck JP, Nielsen R (1999) Variation in the pattern of nucleotide substitution across sites. Journal of Molecular Evolution 48:86-93

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310-2314

Hughes AL (1997) Evolution of the proteasome components. Immunogenetics 46:82-92

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Computer Application in Bioscience 8:275-282

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, New York, USA, p 21-132

Kelsey CR, Crandall KA, Voevodin AF (1999) Different models, different trees: the geographic origin of PTLV-I. Molecular Phylogenetics and Evolution 13:336-347

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624-626

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16:111-120

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proceedings of the National Academy of Sciences, USA 78:454-458

Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. Journal of Molecular Evolution 31:151-160

Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution 11:459-468

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: Molecular Evolutionary Genetics Analysis software. Arizona State University, Tempe

Larget B, Simon D (1999) Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. Molecular Biology and Evolution 16:750-759

Lemmon AR, Moriarity EC (2004) The importance of proper model assumption in Bayesian Phylogenetics. Systematic Biology 53:265-277

Li W-H (1997) Molecular Evolution. Sinauer, Sunderland, MA

Merritt TJS, Quattro JM (2001) Evidence for a period of directional selection following gene duplication in a neutrally expressed locus of Triosephosphate Isomerase. Genetics 159:689-697

Moore RC, Purugganan MD (2003) The early stages of duplicate gene evolution. Proceedings of the National Academy of Sciences USA 100:15682-15687

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Molecular Biology and Evolution 11:715-724

Navidi WC, Churchill GA, von Haeshler A (1991) Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Molecular Biology and Evolution 8:128-143

Nei M (1996) Phylogenetic analysis in molecular evolutionary genetics. Annual Review of Genetics 30:371-403

Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proceedings of the National Academy of Sciences, USA 94:7799-7806

Page RDM, Holmes EC (1998) Molecular Evolution: A Phylogenetic Approach. Blackwell Science, Cambridge

Pedersen A-MK, Wiuf C, Christiansen FB (1998) A codon-based model designed to describe lentiviral evolution. Molecular Biology and Evolution 15:1069-1081

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818

Posada D, Crandall KA (2001a) Selecting the best-fit model of nucleotide substitution. Systematic Biology 50:580-601

Posada D, Crandall KA (2001b) Intraspecific gene genealogies: trees grafting into networks. Trends in Ecology and Evolution 16:37-45

Posada D, Crandall KA (2002) The effect of recombination on the accuracy of phylogeny estimation. Journal of Molecular Evolution 54:396-402

Raftery AE (1996) Hypothesis testing and model selection. In: Gilks WR, Richardson S, Spiegelhalter DJ (eds) Markov chain Monte Carlo in practice. Chapman & Hall, London, p 163-187

Richards MH, Nelson JL (2000) The evolution of vertebrate antigen receptors: a phylogenetic approach. Molecular Biology and Evolution 17:146-155

Rodriguez F, Oliver JF, Marin A, Medina JR (1990) The general stochastic model of nucleotide substitution. Journal of Theoretical Biology 142:485-501

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574

Rzhetsky A, Nei M (1995) Tests of applicability of several substitution models for DNA sequence data. Molecular Biology and Evolution 12:131-151

Rzhetsky A, Sitnikova T (1996) When is it safe to use an oversimplified substitution model in tree making? Molecular Biology and Evolution 13:1255-1265

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425

Sanderson MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. Molecular Biology and Evolution 14:1218-1232

Sanderson MJ, Kim J (2000) Parametric phylogenetics? Systematic Biology 49:817-829

Satta Y, Kupferman H, Li Y-J, Takahata N (1999) Molecular clock and recombination in primate MHC genes. Immunological reviews 167:367-379

Schierup MH, Hein J (2000) Consequences of recombination on traditional phylogenetic analysis. Genetics 156:879-891

Schierup MH, Mikkelsen AM, Hein J (2001) Recombination, balancing selection, and phylogenies in MHC and self-incompatibility genes. Genetics 159:1833-1844

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6:461-464

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular Biology and Evolution 16:1114-1116

Simmons MP, Pickett KM, Miya M (2004) How Meaningful Are Bayesian Support Values? Molecular Biology and Evolution 21:188-199

Sorhannus U, Van Bell C (1999) Testing for equality of molecular evolutionary rates: a comparison between a relative-rate test and a likelihood ratio test. Molecular Biology and Evolution 16:849-855

Suchard MA, Weiss RE, Sinsheimer JS (2001) Bayesian Selection of Continuous-Time Markov Chain Evolutionary Models. Molecular Biology and Evolution 18:1001-1013

Sullivan J, Holsinger KA, Simon C (1995) Among site rate variation and phylogenetic analysis of 12s rRNA in Sigmontine rodents. Molecular Biology and Evolution 12:988-1001

Sullivan J, Swofford DL (2001) Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Systematic Biology 50:723-729

Suzuki Y, Glazko GV, Nei M (2002) Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. PNAS 99:16138-16143

Swofford DL (1998) PAUP* Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.0. Sinauer, Sunderland, MA

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular Systematics. Sinauer, Sunderland, MA

Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Systematic Biology 50:525-539

Tajima F, Takezaki N (1994) Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Molecular Biology and Evolution 11:278-286

Takahashi K, Nei M (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Molecular Biology and Evolution 17:1251-1258

Takezaki N, Zaleska-Rutczynska Z, Figueroa F (2002) Sequencing of amphioxus *PSMB5/8* gene and phylogenetic position of agnathan sequences. Gene 282:179-187

Tamura K (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. Molecular Biology and Evolution 9:678-687

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in Humans and Chimpanzees. Molecular Biology and Evolution 10:512-526

Tavare S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lec. Math. Life Sci. 17:57-86

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673-4680

Wakeley J (1994) Substitution rate variation among sites and the estimation of transition bias. Molecular Biology and Evolution 11:436-442

Whelan S, Goldman N (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Molecular Biology and Evolution 16:1292-1299

Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular Biology and Evolution 18:691-699

Whelan S, Lio P, Goldman N (2001) Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends in Genetics 17:262-272

Wilcox TP, Zwickl DJ, Heath TA, Hillis DM (2002) Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Molecular Phylogenetics and Evolution 25:361-371

Yang Z (1993) Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution 10:1396-1401

Yang Z (1994a) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39:306-314

Yang Z (1994b) Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Systematic Biology 43:329-342

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology and Evolution 11:367-372

Yang Z, Goldman N, Friday A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Molecular Biology and Evolution 11:316-324

Yang Z, Goldman N, Friday A (1995) Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. Systematic Biology 44:384-399

Yang Z, Nielsen R, Goldman N, Pedersen A-MK (2000) Codon substitution models for heterogeneous selection pressure and amino acid sites. Genetics 155:431-449

Yoder AD, Yang Z (2000) Estimation of primate speciation dates using local molecular clocks. Molecular Biology and Evolution 17:1081-1090

Zhang J, Gu X (1998) Correlation between the substitution rate and rate variation among sites in protein evolution. Genetics 149:1615-1625

Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. Journal of Molecular Evolution 39:315-329

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving Genes and Proteins. Academic Press, New York, p 97-166

CHAPTER II

NATURAL SELECTION DURING FUNCTIONAL DIVERGENCE OF *LMP7* AND PROTEASOME SUBUNIT X (*PSMB5*) FOLLOWING GENE DUPLICATION

ABSTRACT

The *lmp7* and *psmb5* genes were created through an ancient gene duplication event of their ancestral locus. These proteins contain an active site of proteolysis, and LMP7 replaces PSMB5 as a component of the 20S proteasome after stimulation of cells by interferon-γ. Replacement of PSMB5 by LMP7 changes the profile of the products of 20S proteasome processing, predisposing digested peptides for transport to and display by the immune system. The purpose of this chapter is to investigate evolutionary forces influencing functional divergence between *lmp7* and *psmb5* following duplication. Levels of synonymous and nonsynonymous substitution rates are estimated to infer differences in levels of natural selection. Estimates of substitution rates indicate that natural selection elevated rate of nonsynonymous substitution in *lmp7* following gene duplication, whereas *psmb5* experienced an increase in substitution rate that was not likely due to diversifying natural selection following duplication. Following initial divergence, nearly neutral mutations have dominated gene evolution in both lineages. The *lmp7* gene locus provides a rare example of a protein with specialized function arising from duplication and divergence of a housekeeping protein by way of natural selection.

INTRODUCTION

Gene duplications are an important factor in molecular evolution and are a major mechanism through which proteins can assume new or specialized functions. Through duplications of genomes, chromosomes or genes, several copies of a gene may arise and later diverge to form multigene families (Li 1997; Ohno 1970). Duplication events can result in one of several outcomes, brought about by a combination of various competing mechanisms. For the vast majority of gene duplications, redundant gene loci will likely degenerate into nonfunctional pseudogenes due to the deleterious nature of most mutations and the initial low frequency of the haplotype (Lynch and Conery 2000; Walsh 1995). However, models of duplicate gene loss are difficult to reconcile with the relatively high numbers of duplicate loci found in the genomes of some model organisms (Hughes and Hughes 1993; Prince and Pickett 2002). Recently, theoretical work has focused on mechanisms that preserve duplicated genes from loss due to a null mutation.

The result of a gene duplication can be influenced by several evolutionary factors. When an ancestral gene has multiple roles, subfunctionalization theory predicts that both gene copies may be preserved and each assumes a different subset of ancestral functions (Force et al. 1999). Under this scenario, a gene copy may become unable to perform particular ancestral functions due to the deterioration of one or more regulatory regions, and the other gene copy is then preserved to carry out these functions. Gene duplicates can also be preserved due to purifying selection that occurs because the protein has multiple domains or is part of a molecule with several subunits (Gibson and Spring 1998). In this case, point mutations may result in a stronger phenotype than a null mutation, and as long as gene duplicates are expressed, it is possible to be preserved as a subunit of a molecule due to purifying selection on the structure of the multiple domains or multi-subunit molecule. Finally, it has been proposed that a gene copy can persist long enough to specialize or acquire a new function through the forces of positive (diversifying) natural selection (Hill and Hastie 1987). Divergence according to positive natural selection is often manifested through higher rates of amino acid substitution when compared to the synonymous substitution rate in DNA. In fact, all the mechanisms of gene locus maintenance outlined above often involve a change in the rate of substitutions from the basal rate.

The period after a gene duplication is often marked by an interval of increased substitutions but the cause of this increased rate is debated. Initially, the rate increase was attributed to neutral mutations due to relaxed selective constraints on duplicate gene copies (Ohno 1970). More recently, a role for natural selection at the molecular level has been postulated (Hughes 1994). However, it is often problematical to distinguish the effects of neutral evolution and evolution by natural selection. Furthermore, natural selection may be difficult to demonstrate because it is likely to occur at only a few sites in a sequence and may exert influence for only a short amount of time and act differently on gene duplicates (Golding and Dean 1998; Yang 2001). Despite the difficulty characterizing older duplication events, they are of particular interest because of the high numbers of duplication events reported to occur in early vertebrate evolution (Gu et al. 2002; McLysaght et al. 2002; Ohno 1970). Some loci involved in duplications during early vertebrate evolution include genes of the adaptive immune system, many of which have related duplicates in paralogous regions of the genome (Abi Rached et al. 2002; Flajnik and Kasahara 2001).

The 20S proteasome is a vital housekeeping component of the cell and is responsible for the constant degradation of cellular proteins into short peptides and amino acids (Rock et al. 1994). The 20S proteasome is comprised of α and β subunits, of which some β subunits contain the active site of proteolysis (Arendt and Hochstrasser 1997; Heinemeyer et al. 1997; Seemuller et al. 1995). In most vertebrates, stimulation of cells by interferon-γ alters the biochemical profile of cleavage sites and size spectrum of peptide products (Boes et al. 1994; Driscoll et al. 1993), a result of the replacement of the three conventionally expressed active β subunits by closely related gene family members. The conventionally expressed subunits of the housekeeping proteasome, X, Y and Z (human genome database coded as PSMB5, PSMB6 and PSMB7 respectively), are replaced by LMP7 (PSMB8), LMP2 (PSMB9) and MECL1 (PSMB10) respectively (Coux 1996). When these three new subunits are in place, the action of the proteasome changes so that proteins are cleaved more frequently after hydrophobic residues and less after acidic residues (Gaczynska et al. 1994; Toes et al. 2001). These peptides are more effectively transported to the endoplasmic reticulum (ER) and loaded onto major histocompatibility complex (MHC) class I proteins that are displayed on the cell surface to cytotoxic T-cells (Coux 1996; Rock et al. 2002).

Proteasome components LMP7 and LMP2 are encoded in the MHC regions of vertebrates along with the MHC class I genes and other functionally related genes of the immune system (Flajnik and Kasahara 2001). Proteasomes with these interferon-γ inducible subunits are called immunoproteasomes, and have already been shown to be functionally distinct from the constitutively expressed proteasomes due to the incorporation of active β subunits LMP2, LMP7 and MECL1 (Boes et al. 1994; Kesmir et al. 2003). The constitutive and interferon-γ inducible proteasome subunit pairs originate from duplication of the three ancestral loci (Hughes 1997). Linkage patterns and inferred homology has indicated that the three interferon-γ inducible forms were possibly created by simultaneous chromosomal duplication of the more ancient *psmb5, 6* and *7* (Clark et al. 2000; Kasahara et al. 1996).

The mechanism of functional diversification since duplication is of particular interest, and two main classes of theories on the predominant form of diversification have been proposed. The main difference between paradigms of diversification is the emphasis placed on neutral mutation and drift versus natural selection (Wagner 2002; Zhang et al. 1998b). Previous work on *psmb5* and *lmp7* found evidence for an increased substitution rate in interferon-γ inducible proteasome subunits, possibly associated with acquisition of specialized (sub)function (Takezaki et al. 2002). Here, I build on the work of Takezaki et al. (2002) by further investigating the nature and cause of the elevated substitution rates in specific lineages.

Examples of positive selection operating to diversify the function of vertebrate gene families are common is the current literature, but are mostly limited to loci involved in reproductive isolation or non-self recognition. In this research, the focus is to study sequences of *psmb5* and *lmp7* to elucidate molecular evolutionary forces causing nucleotide divergence. I specifically target the interval following the duplication of the proto-*psmb5/lmp7* housekeeping gene, and investigate substitution rates to infer the operation of diversifying natural selection. Results indicate that this is a rare example of divergence of an essential housekeeping gene (the proto-*psmb5/lmp7*) involving duplication and evolution under natural selection.

MATERIALS AND METHODS

DATA

Twenty-one sequences were downloaded from the Genbank database. These data are from a variety of vertebrate taxa and are comprised of *lmp7* and/or *psmb5* coding regions. Sequences are primarily from vertebrate taxa because interest is in the evolution of the paralogs created after gene duplication. Sequence from a tunicate and a lancet were included in the data to isolate the period of evolutionary time prior to and following the putative duplication event and because evidence suggests that these organisms are close outgroups to *psmb5* and *lmp7* (Takezaki et al. 2002). All sequences were aligned in Clustal X (Thompson et al. 1994) and DNA sequences were checked to ensure that the alignment preserved the coding frame. The resulting alignment consists of 585 nucleotides and is the same as that found by Takezaki et al. (2002). Testing for saturation was done using the index of substitution saturation (Xia et al. 2003). The index score (I_{SS}) is significantly smaller than the critical score ($I_{SS.C}$) for these data ($I_{SS} = 0.49$; $I_{SS.C} = 0.72$; $P < 0.001$).

SUBSTITUTION RATE ESTIMATION

Models of codon evolution that estimate numbers of nonsynonymous (d_n) and synonymous (d_s) substitutions, and calculate $d_n/d_s = \omega$ rates are used to infer levels of natural selection (Nielsen and Yang 1998; Yang et al. 2000). Overall, eleven models are included as candidates to describe these data, and codon frequencies are estimated using the F3X4 option. Models are designated as being site-specific, branch-specific, or branch-site models. A simple one-rate model is described in Goldman and Yang (1994) and is included for direct comparison to other models and the possibility that the data cannot support inference from more complex models. Site-specific models are described in Yang et al. (2000) and branch-specific and branch-site models are laid out in Yang (1998) and Yang and Nielson (2002) respectively. For branch-specific and branch-site models, evolutionary lineages with the basal substitution rate are referred to as being in the "background" or having the background rate, and lineages targeted as having a different rate are said to be in the "foreground" of the tree topology (see Table 2.1 and Figure 2.1).

The same branching topology reconstructed by Takezaki et al. (2002) is used here, and it indicates that duplication of ancestral proteasome components occurred after divergence of vertebrates from amphioxus and prior to the separation of gnathostomes and

agnathans. Because substitution rates are not constant in this tree, I primarily use models that combine rate variation both in time and among codons to more accurately approximate these findings (Burnham and Anderson 2002). Variants of branch-specific or branch-site models focus on lineages in which natural selection may have operated for a short time. These branches include the ancestral lineage to all vertebrates, lineages immediately following the duplication event (including the ancestral lineage of *psmb5* in jawed vertebrates), and the common ancestral lineage of *lmp7*. These lineages are *a priori* designated as foreground branches in various models because substitutions that occur shortly after a duplication or lineage divergence can have a large impact on the subsequent evolution of gene loci, and the relative forces directing these substitutions are of interest.

Model parameters were estimated using CODEML in the PAML package (Yang 2000) , and empirical Bayes methods are used to identify which specific sites are likely to

Model	fp[1]	rates1[2]	rates2[3]	foreground	parameters
one rate model					
M0	50	1	1	none	ω_0
site-specific models					
M3 (k = 3)	54	1	3	none	$p_0\ p_1\ p_2$ $\omega_0\ \omega_1\ \omega_2$
M3 (k = 2)	52	1	2	none	$p_0\ p_1$ $\omega_0\ \omega_1$
branch-specific models (two-ratios)					
Br1	51	2	1	branch 1	$\omega_0\ \omega_1$
Br2	51	2	1	branch 2	$\omega_0\ \omega_1$
Br3	51	2	1	branch 3	$\omega_0\ \omega_1$
Br4	51	3	1	branch 4	$\omega_0\ \omega_1$
branch-site models (model B of Yang and Nielson (2002))					
MB-1	54	2	3	branch 1	$p_0\ p_1\ p_2\ p_3$ $\omega_0\ \omega_1\ \omega_2$
MB-2	54	2	3	branch 2	$p_0\ p_1\ p_2\ p_3$ $\omega_0\ \omega_1\ \omega_2$
MB-3	54	2	3	branch 3	$p_0\ p_1\ p_2\ p_3$ $\omega_0\ \omega_1\ \omega_2$
MB-4	54	2	3	branch 4	$p_0\ p_1\ p_2\ p_3$ $\omega_0\ \omega_1\ \omega_2$

[1] number of free parameters including branch lengths
[2] number of rates among branches
[3] number of rates among codon sites

Table 2.1. Candidate models used to estimate codon substitution rates. Models vary by having a number of different classes of substitution rates among codon sites, branches, or both.

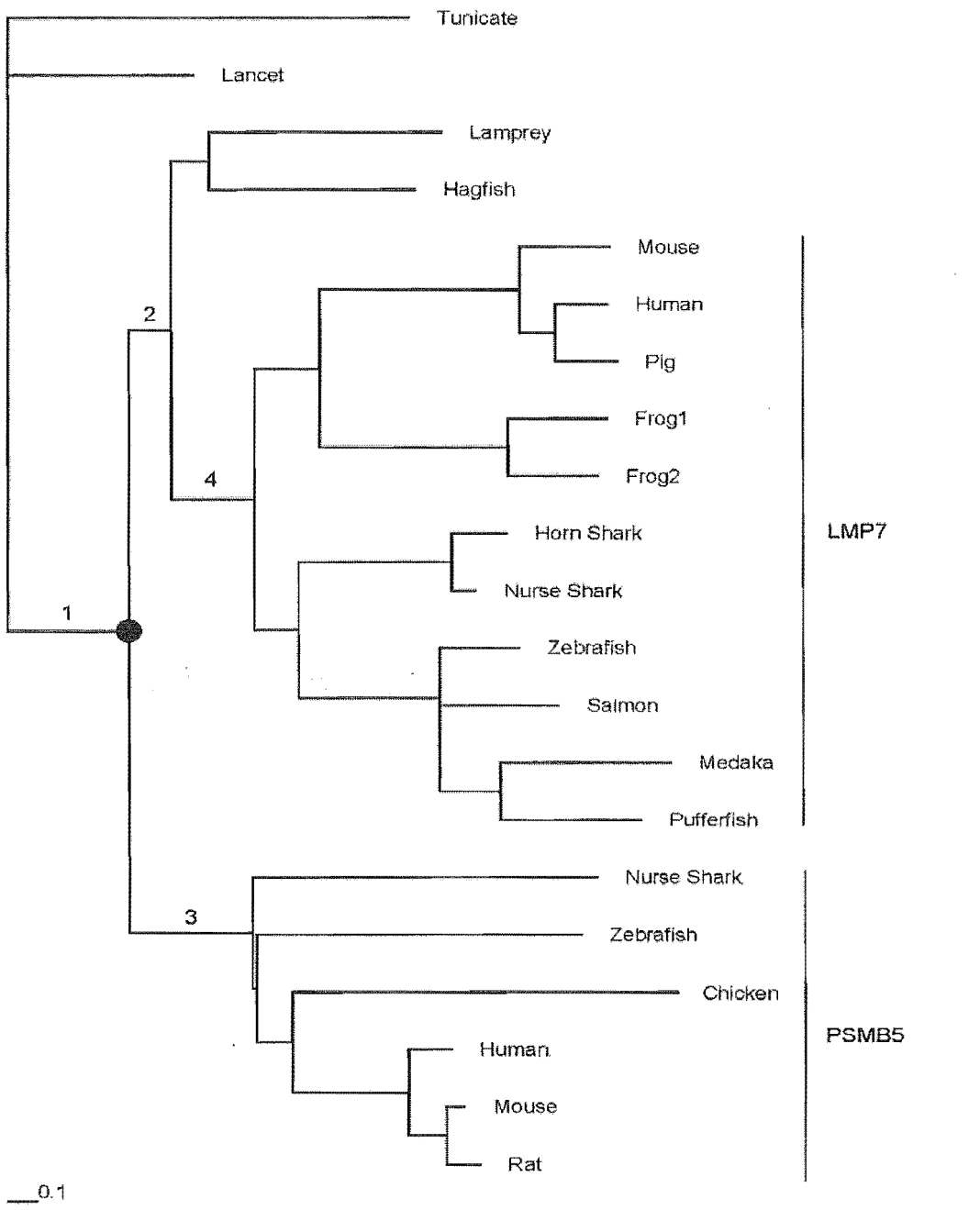


Figure 2.1. Evolutionary relationships among taxa inferred from Takezaki et al (2002) using *LMP7* and *PSMB5* sequences. Branch lengths shown are from the best approximating model derived from current analyses. The putative duplication event is shown as a closed circle and numbered branches are those used as foreground branches. Accession numbers are: AF449497 (lancet, *Branchiostoma lanceolatum*); X97729 (tunicate, *Botryllus schlosseri*); D64054 (hagfish, *Myxine glutinosa*); D64055 (lamprey, *Petromyzon merinus*); *PSMB5*: D29011 (human, *Homo sapiens*); AF060091 (mouse, *Mus musculus*); D45247 (rat, *Rattus rattus*); AB001935 (chicken, *Gallus gallus*); AF155578 (zebrafish, *Danio rerio*); D64058 (nurse shark, (*Ginglymostoma cirratum*); *LMP7*: BC001114 (human, *Homo sapiens*); AF059493 (pig, *Sus scrofa*); U22032 (mouse, *Mus musculus*); D44549, D44540 (African clawed frog, *Xenopus laevis*); D89725 (medaka, *Orizyias latipes*); AJ271723 (pufferfish, *Fugu rubripes*); AF184938 (salmon, *Salmo salar*); AF032390 (zebrafish, *Danio rerio*); D64057 (nurse shark, *Ginglymostoma cirratum*); AF363583 (horn shark, *Heterodontus franscisi*).

fall into various substitution rate categories, thus identifying sites with $\omega > 1$. Maximum likelihood estimation procedures are used because they have been shown to perform well for deeply divergent genes (Muse 1996). These methods have been used on older duplication events (Rodriguez-Trelles et al. 2003), and similar methods have been shown to be accurate at comparable divergence levels (Anisimova et al. 2001). Further, these data do not show signs of substitution saturation so that estimates of substitution rates should be reliable. It is also noted that estimating substitution rates of divergent genes may result in underestimation of substitutions, reducing the difference between d_n and d_s, making inference of selection more conservative (Suzuki and Nei 2001). Akaike's information criterion (AIC) (1974) is used to rank candidate models according to how well the model describes the patterns of substitutions in the data. Since AIC scores are relative measures, the model with the smallest score is taken as the benchmark, and the difference between the best score and all other candidate model scores (δAIC) is presented. Once models are ranked based on ability to describe the data, inference is drawn from parameter estimates of the best approximating model.

RESULTS

MODEL FITNESS AND SELECTION

Maximum likelihood scores of the eleven candidate models varied depending on which factors were included. Generally, models that do not account for rate variation among sites fit the data more poorly than models that include parameters for variation among codons (Table 2.2). Likelihood scores of models that account for rate variation among both lineages and codons were much higher than those without among-site codon rate variation. Among branch-specific and branch-site models, those having branch 4 in the foreground had the highest likelihood values. Overall, MB-4 had the highest likelihood and besides M3(k = 3), other models had much lower likelihood scores (Table 2.2). Likelihood values will always be better for more complex models, so AIC statistics are calculated from likelihood scores in order to more directly compare models relative to each other and estimate how well they approximate the data while taking into account model complexity.

Candidate models differed widely in δAIC scores, although a few models clustered closely together with similar scores (Table 2.2). Typically, models with δAIC scores < 2 are considered to have substantial support as a good approximation to the data and models

model	lnL	δAIC	parameter values
MB-4	-6718.731	best	$p_0 = 0.242$ $p_1 = 0.586$ $p_2 = 0.050$ $p_3 = 0.122$ $\omega_0 = 0.186$ $\omega_1 = 0.013$ $\mathbf{\omega_2 = 999}$
M3 (k = 3)	-6720.760	4.06	$p_0 = 0.621$ $p_1 = 0.293$ $p_2 = 0.086$ $\omega_0 = 0.010$ $\omega_1 = 0.108$ $\omega_2 = 0.347$
MB-1	-6730.537	23.61	$p_0 = 0.014$ $p_1 = 0.035$ $p_2 = 0.278$ $p_3 = 0.673$ $\omega_0 = 0.184$ $\omega_1 = 0.015$ $\omega_2 = 0.00$
M3 (k = 2)	-6739.573	37.68	$p_0 = 0.700$ $p_1 = 0.300$ $\omega_0 = 0.014$ $\omega_1 = 0.183$
MB-3	-6738.075	38.69	$p_0 = 0.278$ $p_1 = 0.659$ $p_2 = 0.019$ $p_3 = 0.045$ $\omega_0 = 0.186$ $\omega_1 = 0.014$ $\mathbf{\omega_2 = 3.34}$
MB-2	-6739.482	41.50	$p_0 = 0.164$ $p_1 = 0.383$ $p_2 = 0.136$ $p_3 = 0.317$ $\omega_0 = 0.183$ $\omega_1 = 0.014$ $\omega_2 = 0.00$
Br4	-6940.957	438.45	$\omega_0 = 0.053$ $\mathbf{\omega_1 = 999}$
Br1	-6950.375	457.29	$\omega_0 = 0.054$ $\omega_1 = 0.000$
M0	-6953.428	461.39	$\omega_0 = 0.055$
Br3	-6952.639	461.82	$\omega_0 = 0.055$ $\omega_1 = 0.206$
Br2	-6953.215	462.97	$\omega_0 = 0.055$ $\omega_1 = 0.313$

Table 2.2. Results of ML optimization of substitution parameters and δAIC rank. Proportions of codon sites (p_i) falling into a particular substitution rate (ω_i) are given. Substitution rates greater than 1 are boldface font, and are taken to signify the operation of natural selection.

with δAIC values > 10 have essentially no support from the data (Burnham and Anderson 2002). Among all candidate models, MB-4 was the best approximation to the data. Remaining branch-site models with other lineages in the foreground did not approximate relevant patterns of variation in the data as well and had larger δAIC scores. M3(k = 3) had a small δAIC value, and was the second best approximation overall. Other branch-site and site-specific models have δAIC scores that are greater than 10 and have essentially no empirical support as a good approximation to the data. The δAICs of branch-specific models were more than an order of magnitude larger compared to models with variation among codons. The large differences seen in δAIC scores were also a feature of parameters estimated from different models.

SUBSTITUTION RATIOS

Parameter estimates varied widely with the model used. The best approximating model allows for rate variation among sites and a rate shift in the common ancestral lineage to

homologous copies of *lmp7* (branch 4). According to this model, the majority of sites (58%) are conserved in all branches of the tree, and 24% of sites are moderately conserved, indicating that rates vary across sites for these data. Approximately 17% of sites experienced a temporary shift that elevated substitutions above the basal rate in branch 4, and the d_n/d_s value is greater than one in this lineage (Tables 2.2 and 2.3). There are several substitutions on branch 4, and in the analysis of Takezaki et al. (2002), this lineage is strongly supported by a high bootstrap value. Despite the abundant numbers of total substitutions, too few synonymous changes are inferred on this branch to accurately estimate ω_2. Substitution rates that result from using other models also denote variation in substitution rates within lineages and among codon sites.

$0.5 > p \geq 0.01$	$0.01 > p$
30	21
58	24
62	32
67	46
88	48
107	53
136	65
179	77
186	84
	87
	91
	99
	114
	125
	129

Table 2.3. Codon sites with $\omega > 1$ under model MB-4. These sites are from the *lmp7* gene locus and rate changes occur shortly after duplication of the proto-*lmp7* locus.

Models with lineages other than branch 4 in the foreground also indicated rate shifts. For all branch-site models, background branches were conserved, the majority of sites with $\omega < 0.02$. Both branches 3 and 4 had $\omega_2 > 1$ estimates in the foreground for sites that changes evolutionary rate. Other foreground branches had a decrease in d_n/d_s for sites that may have temporarily changed rates (Table 2.2). Site-specific models also estimated a variation in substitution rate among codons, with most being highly conserved and none with $\omega > 1$. Since these models average substitution rates across all lineages, a temporary elevation in nonsynonymous substitution rate would likely not result in $\omega > 1$. Nevertheless, variation in rates among codons is a relatively important aspect of evolution in these proteins, as site-specific models were a better fit to the data than branch-specific models. In addition, a direct comparison between each branch-site model and M3(k = 2) is possible because these models are nested. δAIC score comparison reveals that adding a rate shift in branch 4 represents a substantial improvement in the fit of the model, whereas a rate shift in other branches tested here does not result in an improvement of the model.

DISCUSSION

MODEL FITNESS

The set of candidate models used here to investigate *psmb5/lmp7* divergence since duplication allow the investigation of the relative roles of competing evolutionary forces through patterns of nucleotide substitution. The best approximating model indicates that there is rate variation among codons and also a substantial rate shift in branch 4 of the tree topology. This shift results in an elevated nonsynonymous substation rate that is most easily explained by the operation of natural selection for a short period of time. Differences in population size or effectiveness of purifying selection may also lead to elevated nonsynonymous substitution rate, but overlapping taxonomic sampling between *lmp7* and *psmb5* make this explanation less likely. Results also indicate an increase in the rate of evolution in branch 3, but this model did not represent the data better than models with no temporary rate shift. Therefore, the evidence of natural selection in this branch is unconvincing because it does not improve the fit of the model. The ancestral lineage to *psmb5* homologues probably did experience a rate shift, but it is not clear that it was due to natural selection. Subsequently, a model was used that included both branches 3 and 4 in the foreground, but this models did not have an improved fit to the data (*ln*L -6738.726) compared to site-specific models, probably due to different sites with elevated rates of evolution in respective branches. Inference from parameters estimated from the best approximating model is informative to identify prevalent forces acting to diversify duplicate gene loci.

PERSISTENCE OF DUPLICATED GENES

Several theories explain the maintenance of duplicated loci that are not mutually exclusive, and the genomic organization and functions of *psmb5* with *lmp7* indicate that multiple factors may have contributed. Immediately following the duplication event, both loci were likely expressed and incorporated into proteasomes, so it is possible that these loci were maintained in the gnathostome lineage due to negative selection as subunits of the proteasome (Gibson and Spring 1998). For duplicated proteasome components, a null mutation might not be devastating to the function of the proteasome because of redundancy, but a locus with a deleterious mutation to a site directly involved in proteolysis could reduce efficiency in a sizeable fraction of the proteasomes in a cell.

Therefore, purifying selection may have played an early role in preserving duplicate proteasome components in most descendant lineages.

Since *psmb5* and *lmp7* currently have different expression patterns in tissue and timing (Akiyama et al. 1994), it is very likely that complementary degenerative mutations in regulatory regions also contributed to persistence of both loci, if not immediately, then shortly after duplication. When different regulatory regions are inactivated, then it is possible that the individual loci assumed distinct ancestral roles, although at this point it is likely that either locus could perform the suite of ancestral roles equally well. Nevertheless, the differences in expression patterns may have helped contribute to promoting differences between the loci, and initiated processes that lead to functional diversification.

FUNCTIONAL DIVERSIFICATION

Results from the best approximating model indicate that nonsynonymous substitution rates varied widely across sites in the sequences. The selective pressure to maintain structure and function differs in various parts of the PSMB5/LMP7 proteins; however, there are several substitutions that are specific to just one subfamily (Hayashi et al. 1997; Kandil et al. 1996). Although agnathan sequences cluster with gnathostome *lmp7*, they lack LMP7-specific substitutions near the active site. Of particular interest is the cassette of codons in positions 27-31 that are near the active site and have radical substitutions in *lmp7* but not *psmb5*. These sites may be identified evolving under natural selection for a time or as being "constant but different" sites that reflect differing roles of structure and function among the two subfamilies (Gribaldo et al. 2003).

In addition to the active site, the S1 pocket plays a crucial part in lysis by β subunits (Lowe et al. 1995). In PSMB5/LMP7 this pocket plays a key role in specificity of cleavage due to steric interactions and biochemical properties of the pocket (Toes et al. 2001). In PSMB5/LMP7 homologues, the S1 pocket is comprised of amino acids from two adjacent subunits, and concerted movement and rotation of side chains in and near the S1 pocket in conjunction with substrate contact has been documented, (Groll et al. 1997). Due in part to these complexities, the exact mechanisms responsible for functional divergence between constitutive and interferon-γ inducible forms of this β subunit remain unclear. Nevertheless, crystal structures have helped identify amino acids that comprise the reactive core (1, 17, 33), bind to the active-site residues (129, 166, 168), form and

determine the character of the S1 pocket (20, 31, 35, 45, 49, 53), and comprise additional residues in contact with substrate undergoing lysis in the proteasome (21, 47) (Groll et al. 1997; Unno et al. 2002).

Twenty-four sites are identified in these data as evolving for some time under natural selection (Table 2.3). Several of the selected residues in Table 2.3 are or are adjacent to residues identified above as involved in function and specificity of LMP7. Substitutions in *lmp7* near the above residues are positions 21, 30, 32, 46, 48, 125 and 129, and some of these sites also involve radical changes in amino acid properties. The positioning of these substitutions makes it possible that these sites have been involved with the functional divergence of these subfamilies by altering the stereochemistry of the S1 pocket. These changes may also alter the capacity for steric changes that occur in concert with substrate binding, or the stoichiometry of the region surrounding the S1 pocket. Many substitutions taking place by natural selection occurred shortly after gene duplication, in branches ancestral to vertebrate *lmp7* homologues.

Branches 3 and 4 represent the ancestral lineage to *psmb5* and *lmp7* respectively and have a higher rate of evolution than the basal rate found for the entire tree. The ω_2 estimate for branches 3 and 4 are both above 1. However, only an elevated rate of evolution in branch 4 improves model fitness, an indication that natural selection was acting for a time to diversify ancient lineages of the *lmp7* locus, but the elevated substitution rate at the *psmb5* locus was more likely due to relaxed selective constraints. The operation of natural selection in the ancestral branch of a functionally divergent gene locus supports the hypothesis that LMP7 diverged in function through positively selected amino acid substitutions. Since that time however, it is likely that nearly neutral evolution has dominated substitution rates in *lmp7*. The *lmp7* gene locus provides an example of a protein with specialized function arising from duplication, divergence and natural selection of a housekeeping protein.

ACKNOWLEDGEMENTS

Bruce Waldman, Neil Gemmel, Martin Flajnik, and peer referees provided valuable suggestions on this manuscript. Ziheng Yang is gratefully acknowledged for support with various versions of PAML. This work is supported by the Marsden Fund of New Zealand and by a University of Canterbury Doctoral Scholarship.

REFERENCES

Abi Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H (2002) Evidence of *en bloc* duplication in vertebrate genomes. Nature Genetics 31:100-105

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control AC 19:716-723

Akiyama K-y, Yokota K-y, Kagawa S, Shimbara N, Tamura T, Akioka H, Nothwang HG, Noda C, Tanaka K, Ichihara A (1994) cDNA cloning and interferon-γ down-regulation of proteasomal subunits X and Y. Science 265:1231-1234

Anisimova M, Beilawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular Biology and Evolution 18:1585-1592

Arendt CS, Hochstrasser M (1997) Identification of the yeast 20S proteasome catalytic centers and subunit interaction required for active-site formation. Proceedings of the National Academy of Sciences, USA 94:7156-7161

Boes B, Hengel H, Ruppert T, Molthaup G, Koszinowski UH, Kloetzel P-M (1994) Interferon-γ stimulation modulates the proteolytic activity and cleavage site preference of 20S mouse proteasomes. Journal of Experimental Medicine 179:901-909

Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference: a practical information-theoretic approach. Springer-Verlag, New York

Clark MS, Pontarotti P, Gilles A, Kelly A, Elgar G (2000) Identification and characterization of a B proteasome subunit cluster in the Japanese Pufferfish (*Fugu rubripes*). Journal of Immunology 165:4446-4452

Coux O (1996) Structure and functions of the 20S and 26S proteasomes. Annual Review of Biochemistry 65:801-847

Driscoll J, Brown MG, Finley D, Monaco JJ (1993) MHC-linked *LMP* gene products specifically alter peptidase activities of the proteasome. Nature 365:262-264

Flajnik MF, Kasahara M (2001) Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. Immunity 15:351-362

Force A, Lynch M, Pickett FB, Amores A, Yan Y-l, Postlethwait JH (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531-1545

Gaczynska M, Rock KL, Spies T, Goldberg AL (1994) Peptidase activities of proteasomes are differentially regulated by the major histocompatibility complex-encoded genes for LMP2 and LMP7. Proceedings of the National Academy of Sciences USA 91:9213-9217

Gibson TJ, Spring J (1998) Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends in Genetics 14:46-49

Golding GB, Dean AM (1998) The structural basis of molecular adaptation. Molecular Biology and Evolution 15:355-369

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11:725-736

Gribaldo S, Casane D, Lopez P, Philippe H (2003) Functional Divergence Prediction from Evolutionary Analysis: A Case Study of Vertebrate Hemoglobin. Molecular Biology and Evolution 20:1754-1759

Groll M, Ditzel L, Lowe J, Stock D, Bochtler M, Bartunik HD, Huber R (1997) Structure of the 20S proteasome from yeast at 2.4A resolution. Nature 386:463-471

Gu X, Wang Y, Gu J (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate gene evolution. Nature Genetics 31:205-209

Hayashi M, Ishibashi T, Tanaka K, Kasahara M (1997) The mouse genes encoding the third pair of *B*-type proteasome subunits regulated reciprocally by INF-γ. Journal of Immunology 159:2760-2770

Heinemeyer W, Fischer M, Krimmer T, Stachon U, Wolf DH (1997) The active sites of the eukaryotic 20 S proteasome and their involvement in subunit precursor processing. Journal of Biological Chemistry 272:25200-25209

Hill RE, Hastie ND (1987) Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature 326:96-99

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proceedings of the Royal Society of London Series B 256:119-124

Hughes AL (1997) Evolution of the proteasome components. Immunogenetics 46:82-92

Hughes MK, Hughes AL (1993) Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. Molecular Biology and Evolution 10:1360-1369

Kandil E, Namikawa C, Nonaka M, Greenberg AS, Flajnik MF, Ishibashi T, Kasahara M (1996) Isolation of low molecular mass polypeptide complimentary DNA clones from primitive vertebrates. Journal of Immunology 156:4225-4253

Kasahara M, Hayashi M, Tanaka K, Inoko H, Sugaya K, Ikemura T, Ishibashi T (1996) Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. Proceedings of the National Academy of Sciences, USA 93:9096-9101

Kesmir C, van Noort V, de Boer RJ, Hogeweg P (2003) Bioinformatic analysis of functional differences between the immunoproteasome and the constitutive proteasome. Immunogenetics 55:437-449

Li W-H (1997) Molecular Evolution. Sinauer, Sunderland, MA

Lowe J, Stock D, Jap B, Zwickl P, Baumeister W, Huber R (1995) Crystal structure of the 20S proteasome from the Archeon *T. acidophilum* at the 3.4 A resolution. Science 268:355-539

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155

McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. Nature Genetics 31:200-204

Muse SV (1996) Estimating synonymous and nonsynonymous substitution rates. Molecular Biology and Evolution 13:105-114

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and application to the HIV-1 envelope gene. Genetics 148:929-936

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, Berlin

Prince VE, Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. Nature Reviews Genetics 3:827-837

Rock KL, Gramm C, Rothstein L, Clark K, Stein R, Dick L, Hwang D, Goldberg AL (1994) Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. Cell 78:761-771

Rock KL, York IA, Saric T, Goldberg AL (2002) Protein degradation and the generaton of MHC class I-presented molecules. In: Dixon FJ (ed) Advances in Immunology. Academic Press, San Diego, CA, p 1-70

Rodriguez-Trelles F, Tarrio R, Ayala FJ (2003) Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the xanthine

dehydrogenase gene. Proceedings of the National Academy of Sciences USA 100:13413-13417

Seemuller E, Lupas A, Stock D, Lowe J, Huber R, Baumeister W (1995) Proteasome from *Thermoplasm acidophilum*: A threonine protease. Science 268:579-582

Suzuki Y, Nei M (2001) Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Molecular Biology and Evolution 18:2179-2185

Takezaki N, Zaleska-Rutczynska Z, Figueroa F (2002) Sequencing of amphioxus *PSMB5/8* gene and phylogenetic position of agnathan sequences. Gene 282:179-187

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673-4680

Toes REM, Nussbaum AK, Degermann S, Schirle M, Emmerich NPN, Kraft M, Laplace C, Zwinderman A, Dick TP, Muller J, Schonfisch B, Schmid C, Fehling H-J, Stevanovic S, Rammensee H-G, Shild H (2001) Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. Journal of Experimental Medicine 194:1-12

Unno M, Mizushima T, Morimoto Y, Tomisugi Y, Tanaka K, Yasuoka N, Tsukihara T (2002) The structure of the mammalian 20S proteasome at 2.75 A resolution. Structure 10:609-618

Wagner A (2002) Selection and gene duplication: a view from the genome. Genome Biology 3:1012.1-1012.3

Walsh JB (1995) How often do duplicated genes evolve new functions? Genetics 139:421-428

Xia X, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. Molecular Phylogenetics and Evolution 26:1-7

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution 15:568-573

Yang Z (2000) Phylogenetic Analysis by Maximum Likelihood (PAML). University College London, London

Yang Z (2001) Adaptive Molecular Evolution. In: Balding DJ, Cannings C, Bishop M (eds) Handbook of Statistical Genetics. John Wiley and Sons, New York, p 327-350

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution 19:908-917

Yang Z, Nielsen R, Goldman N, Pedersen A-MK (2000) Codon substitution models for heterogeneous selection pressure and amino acid sites. Genetics 155:431-449

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proceedings of the National Academy of Sciences, USA 95:3708-3713

CHAPTER III

STRUCTURAL MODELING, NATURAL SELECTION OF MHC PROTEINS AND SUPPORT FOR CO-EVOLUTION AMONG CLASS I REGION GENES IN *XENOPUS*

ABSTRACT

In the African clawed frog (*Xenopus laevis*), two deeply divergent allelic lineages among multiple genes in the MHC class I region have been discovered, namely *lmp7*, *tap1*, *tap2* and class Ia. For the class Ia locus, functional differences and the molecular basis for lineages maintenance are unknown. Alleles of linked class I region genes also exhibit strong linkage disequilibrium with specific class Ia alleles, but the underlying cause is not clear. I use MHC class Ia sequence data to estimate substitution rates and investigate structural differences between allelic lineages from protein models. Results indicate the operation of natural selection, and that alleles can be distinguished based on structure and polymorphism in amino acids of the F pocket. Variation at this site likely enables allelic lineages to bind very different sets of peptides and to interact differently with MHC chaperones in the endoplasmic reticulum. These results constitute the first evidence of the evolutionary basis for 1) the maintenance of allelic lineages, 2) functional differences among lineages and 3) strong linkage disequilibrium of allelic variants of class I region genes in *X. laevis*.

INTRODUCTION

Class Ia MHC proteins play a critical role in the adaptive immune system by presenting peptide fragments derived primarily from endogenous proteins to $CD8^+$ T cells. These peptides are generated in the cytosol through the action of the 20S proteasome and are transported into the endoplasmic reticulum (ER), where they are bound in the peptide binding region (PBR) of class I proteins. Class I proteins associate closely with several chaperone proteins in the ER. Collectively referred to as the "peptide-loading-complex", proteins such as transporter associated with antigen processing (TAP), tapasin (TPN), calreticulin (CRT) and endoplasmic reticulum protein 57 (ERp57) function to assemble, load and optimize peptide selection on class I molecules (Williams et al. 2002). Because of its active role in the adaptive immune response, MHC class I genes are of considerable interest and have been isolated and studied in a variety of organisms.

The African clawed frog (*Xenopus laevis*) is the second non-mammalian species from which MHC proteins were isolated (Flajnik et al. 1984). In this species, there is a single classical MHC class I locus, and two distinctive allelic lineages (A and B) have been found (Flajnik et al. 1999). These lineages have an ancient origin and are found to be as divergent from each other as are mouse and human MHC alleles. Several other genes associated with MHC/peptide generation, transport and loading have been isolated in *X. laevis* (Namikawa et al. 1995; Nonaka et al. 1997a; Nonaka et al. 1997b; Ohta et al. 1999; Ohta et al. 2003). *Tap1, tap2*, and the proteasome component *lmp7* are physically linked to the *X. laevis* class Ia gene in what appears to be the primordial organization of the MHC region (Nonaka et al. 1997a; Ohta et al. 2002). Similar to the class Ia gene, divergent allelic lineages are also found in these MHC processing genes. In some cases, the active sites contain substitutions that may affect functional properties of these proteins (Ohta et al. 2003). Further, alleles of linked MHC processing genes also consistently exhibit strong linkage disequilibrium with specific MHC class Ia lineages.

The strong linkage disequilibrium among alleles at functionally related loci in the MHC occurs among the class Ia gene and certain members of the peptide loading complex. Two members of this complex (CRT and ERp57) are general ER chaperone proteins, but TAP and TPN proteins play an integral role in the loading of high-affinity peptides to class Ia molecules (Grandea III and Van Kaer 2001). Lineage maintenance in TAP loci and in *lmp7* likely result from interaction during peptide loading of different

lineages of class Ia molecules. For their part, allelic lineages in class Ia alleles are hypothesized to differ functionally in their abilities to bind distinct repertoires of peptides. However, the molecular differences among class Ia lineages have not been resolved in enough detail to support or refute this hypothesis. There is also little understanding on how the diversity among lineages in class Ia molecules may relate to variation in peptide loading complex proteins.

High levels of diversity in class I proteins of *X. laevis* and other species are a hallmark of MHC biology and is thought to be necessary to combat infectious parasites in an evolutionary "arms race" (Hill 1999; Potts and Slev 1995). This diversity is due in part to the operation of natural selection (Hughes and Nei 1988; Parham and Ohta 1996), and elevated nonsynonymous substitution rates in the PBR are frequently used to identify the operation of natural selection acting at the molecular level (Apanius et al. 1997; Hughes and Yeager 1998). However, previous results in *Xenopus* indicated that the nonsynonymous substitution rate (d_n) was not significantly greater than the synonymous substitution rate (d_s) in class Ia alleles, despite the common finding of natural selection acting on MHC genes of other taxa (Flajnik et al. 1999). The lack of evidence for natural selection is surprising because the two MHC class Ia lineages are thought to be maintained by natural selection.

While seemingly straightforward, the estimation of substitution rates and inferences concerning natural selection can be difficult (e.g. Crandall et al. 1999). Several difficulties that influence the estimation of substitution rates have been identified, and these obstacles are intensified with the use of shorter sequences (Ina 1995; Muse 1996; Muse and Gaut 1994; Yang 2001; Yang and Nielsen 2000). Therefore, Flajnik and co-workers (1999) concluded that estimates of $d_n/d_s < 1$ in *Xenopus* MHC class Ia PBR are artefacts that arise from saturation of sites rather than lack of natural selection. In addition, many substitution rate estimators also suffer from a larger problem — they estimate an average d_n and d_s across sites and for the entire time separating the sequences in the data set. This results in underestimation of d_n due to variable nonsynonymous substitution rates (Ina 1995).

Methods to estimate site-wise substitution rates of an entire gene sequence have been devised to circumvent problems with averaging rates and small numbers of sites (Nielsen and Yang 1998; Suzuki and Gojobori 1999). The method of Suzuki and Gojobori (1999) is effective at detecting positive selection but because it is based on parsimony, it

does not correct for multiple substitutions that would be expected with highly divergent sequences. Goldman and Yang (1994) described a model of codon evolution that was later extended to more effectively calculate d_n and d_s under the framework of maximum likelihood (ML) parameter estimation (Nielsen and Yang 1998; Yang and Nielsen 2002; Yang et al. 2000). Maximum likelihood methods have been shown to be more powerful at detecting elevated nonsynonymous substitution rates at just a few codon sites, and they more accurately estimate the numbers of substitutions with highly divergent sequences (Muse 1996; Yang and Nielsen 2000).

To estimate levels of natural selection for *X. laevis* class Ia MHC sequences, I use ML methods based on models of codon evolution to estimate d_n and d_s. I also investigate the nature of divergent allelic lineages and co-evolution among class I region proteins in *X. laevis* by molecular modeling and sequence comparison. Initial reports indicated that differences between lineages were most apparent at intracellular components of the protein (Flajnik et al. 1999). In contrast, more recent work shows that differences in the cytoplasmic and transmembrane regions of the protein have little to do with any potential functional differences between lineages (Ohta et al. 2003). Here I focus on sequences of the α1 and α2 domain because natural selection and functional differences between lineages likely originate in the PBR. Class I MHC proteins typically interact most closely with peptide residues at positions 1 and 2 (P1 and P2 respectively), and the C terminal amino acid (PΩ). Specificity of peptides is usually determined by pockets in the PBR that interact with P2 and PΩ. I model class I protein structure from different lineages and investigate differences of the PBR.

MATERIALS AND METHODS

From the Genbank database, I retrieved ten sequences of the class Ia MHC gene from frogs isolated and sequenced by Flajnik et al. (1999). These data (referred to as "*Xenopus*99") are comprised of samples from the species *X. laevis*, *Rana pipiens*, and a laboratory-bred interspecies hybrid of *X. laevis-X. gilli* (accession numbers AF185579-AF185588).

To increase the number of sequences used for analysis of *X. laevis* MHC, I isolated total RNA from blood samples of *X. laevis,* wild-caught from South Africa, using the TRIzol protocol (Life Technologies). I designed primers to amplify exons 2-4 of the MHC class Ia gene with Primer3 software (Whitehead Institute for Biomedical Research,

Massachusetts Institute of Technology) using *X. laevis* class Ia sequences from the Genbank database (forward primer: 5'-GTCACTCCCTGCGYTAYTAT-3'; reverse primer: 5'-TTTCTCCTTCAGGCTGCTGT-3'). I added 1μL RNA to 50 μL PCR with 1μL enzyme cocktail of H-reverse transcriptase and Platinum *Taq*, and 2X buffer from a Superscript One-Step RT-PCR kit (Invitrogen; 0.20 μM final primer concentration). First strand synthesis was performed at 55° C for 25 min.; immediately after first strand synthesis, PCR was performed using the following protocol: 94° C (0:15 min.), 58° C (0:25 min.), 72° C (1:00 min.) for 35 cycles. Each PCR run was preceded by initial denaturing at 94° C (4:00 min.), and followed by 2:00 min. at 72° C. PCR products were cloned into the pCR 4 TOPO TA plasmid vector (Invitrogen), and recombinant DNA was transformed into TOP 10 competent *Escherichia coli* cells. The *X. laevis* MHC insert was sequenced in both directions using Bigdye v3.1 chemistry and an ABI 3730 automated sequencer; each new sequence was isolated and confirmed from multiple clones. Overall eleven chromosomes sampled yielded eleven new sequences. These data combined with *Xenopus*99 are referred to as *Xenopus*04.

I modelled similarities of class I protein structure from *X. laevis* using SWISSMODEL (Guex and Peitsch 1997; Schwede et al. 2003), which searches crystal structures from the protein data bank (PDB) (Berman et al. 2000) and selects appropriate model templates. Initial models are constructed from templates in the "first approach", and the fit of models was improved using conserved alignment features of classical MHC genes (Kaufman et al. 1994). *Xenopus laevis* alleles G and F represent lineages A and B respectively, and were used as modeling targets. Identity shared between target and template was approximately 45%, which is typically satisfactory for constructing good-quality models (Guex et al. 1999). Despite optimization, an element of uncertainty in molecular modeling is recognized because positions of structures depend on peptide interactions, specific bonding, and adjoining residues which may vary from template to target. However, no crystal structures of *X. laevis* MHC are available and protein modeling has proven to be a useful tool for exploring structural features of structurally homologous proteins. SwissPDBviewer (Deepview) (Guex and Peitsch 1997) was used to compare protein structures among allelic lineages. The program CSU (Contacts of Structural Units) (Sobolev et al. 1999) calculated side chain contacts of peptides in templates, and structural features were inferred to be similar in modelled proteins. Putative PBR residues in *X. laevis* and were assumed to be similar to those found in

humans (Saper et al. 1991) due to the conserved nature of the protein structure (Hashimoto et al. 1999; Kaufman et al. 1994).

Each data set was aligned using Clustal W (Thompson et al. 1994). Positions with gaps in the alignment were excluded because of uncertain homology and because models used to estimate ω do not account for gaps in sequence alignments (Yang 2000). Nucleotide diversity was estimated in *Xenopus* with a sliding window of 50 bp using dnaSP v3.51 (Rozas and Rozas 1999). Wu-Kabat variability (Wu and Kabat 1970) was used to investigate amino acid polymorphism in the α1 and α2 domains. Highly polymorphic amino acid sites are defined as having more than twice the average Wu-Kabat score for all sites in the α1 and α2 domains. Recombination may adversely affect the performance of phylogenetic-based estimators of substitution rates (described below) (Anisimova et al. 2003; Shriner et al. 2003), so the program RDP (Martin and Rybicki 2000) was used to detect sequences that are likely recombinants in the frog sequences. Substitution rate estimates described below require reconstruction of evolutionary relationships. A tree topology was inferred for the data sets using the Neighbor Joining (NJ) algorithm (Saitou and Nei 1987) using genetic distance estimates obtained from ML estimation in PAUP* (Swofford 1998). Distance estimates used the best approximating model for those data according to the Akaike Information Criterion (AIC) (Akaike 1974) implemented in MODELTEST (Posada and Crandall 1998).

*Xenopus*04 (with recombinant alleles removed) was used to estimate substitution rates in frogs, and *Xenopus*99 was also run for comparison to earlier results. Substitution parameters were estimated with the ML criterion using several codon-based models as implemented in the CODEML program of the PAML package (Yang 2000). M0 is the simplest model and it specifies a single ω rate averaged across all sites. M3 allows for different ω rates among sites using K different discrete rate categories that are estimated from the data. M7 allows ω rates to vary among sites, but the parameters of the underlying beta distribution (p and q) are constrained to exclude rates for which d_n is greater than d_s. M8 is similar to M7 except a proportion of sites (p_0) falls into the beta distribution and the remaining proportion of sites (p_1) share a single ω ratio that is unconstrained by the distribution. The AIC was used to select among candidate models, and since it is a relative quantity (Burnham and Anderson 2002), score differences (δAIC) between the best model and all other models are reported for model selection. Complex models were run multiple times and starting values were changed to avoid becoming

trapped in a local optimum of the likelihood landscape (Yang 2000). Readers interested in mathematical details of these models are referred to Yang et al. (2000).

RESULTS

POLYMORPHISM

Eleven new sequences of the *X. laevis* class Ia MHC gene were isolated, and high levels of polymorphism were observed among these data (Figure 3.1). The level of polymorphism is much higher than typically found in MHC sequences of well-studied primates, and mimics more closely the diversity of sequences in salmonid fishes (Shum et al. 2001). Nucleotide polymorphisms are found in 209 of 736 sites without gaps. Most polymorphisms (77%) are shared among at least two alleles.

The levels of nucleotide polymorphism varied substantially among coding domains of the mature protein. The diversity of the α1 domain (nt 1 – 255) is quite high, with 50% polymorphic sites. A sliding window analysis indicates that nucleotide diversity in areas of the α1 domain are near 30% (Figure 3.2). The α2 (nt 256 – 534) and α3 (nt 534 – 781) domains have 27% and 12% polymorphic sites respectively. Figure 3.2 shows the marked differences in levels of diversity among the different domains of the class Ia sequences. The α3 domain displays unusual homogeneity compared with the MHC of other organisms and this observation is consistent with findings of Flajnik et al. (1999). The diversity of the α2 domain is considerably lower than the α1 domain even though these domains are both involved in peptide binding. The 3' end displays the highest levels of diversity in the α1 domain, in contrast to the 3' end of the α2 domain which has the lowest levels of diversity, despite homology between these regions.

Amino acid variability was measured using the Wu-Kabat index (Wu and Kabat 1970) on the α1 and α2 domains of the protein, and results are plotted in Figure 3.3. The average Wu-Kabat score for all sites in the α1 and α2 domains is 2.99. There are 18 highly polymorphic amino acid sites in the domains comprising the PBR, and 11 of those sites are in the α1 domain. Distinctive patters of variation differentiate the α1 and α2 domain structures. Seven highly polymorphic sites are in the α1 helix, while only two such sites are found in the α2 helix. Also, all of the highly polymorphic sites in the α1 helix are oriented to interact with the bound peptide, while those found in the α2 helix interact with the T cell receptor. Nine highly polymorphic sites are found nearly evenly

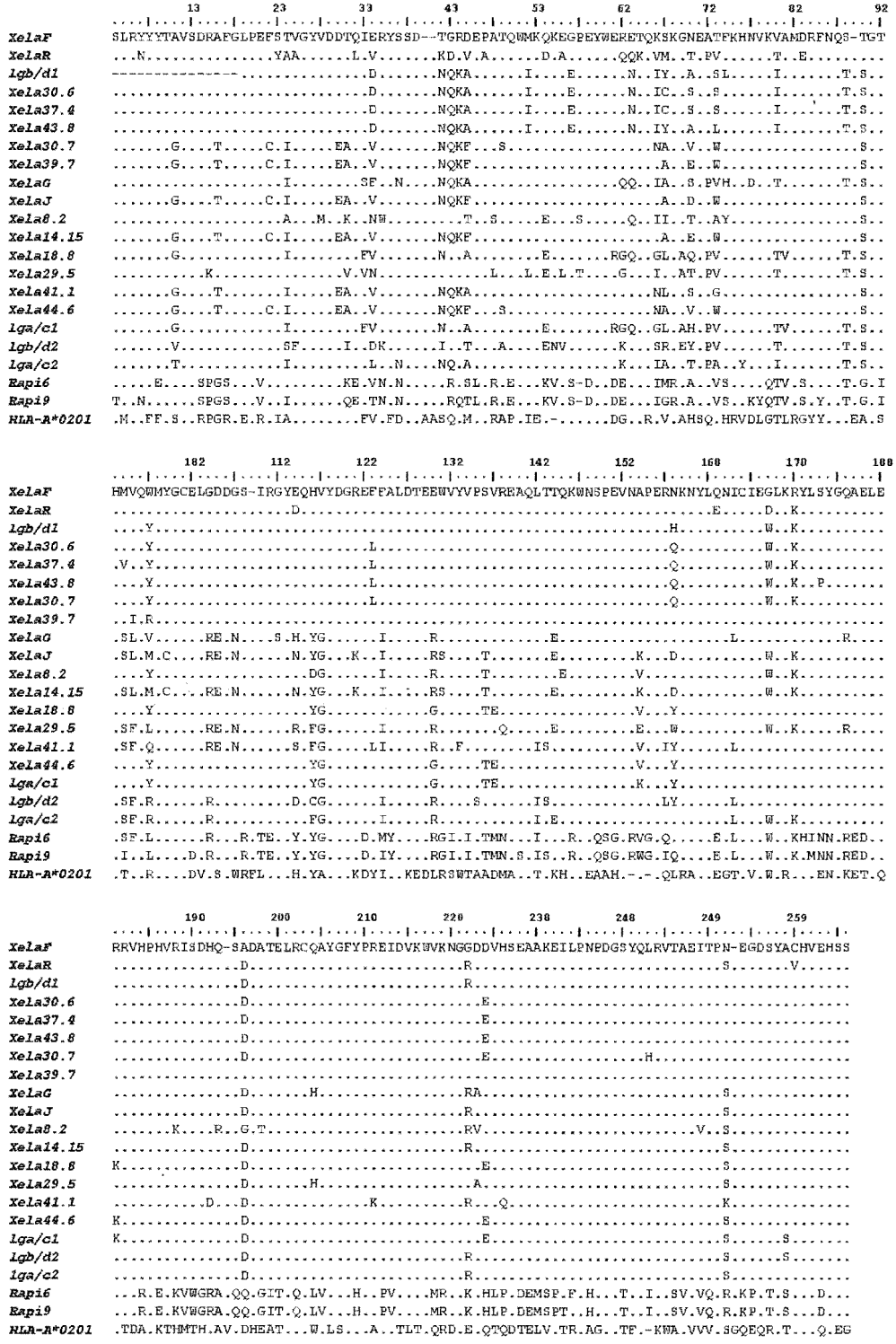


Figure 3.1. Frog MHC class Ia amino acid sequence aligned and numbered according to structural features of HLA-A*0201. HLA sequence and recombinant sequences Xela R, Xela 30.7, Xela 39.7, Xela 14.15, Xela 44.6, and lg b/d2 not included in substitution rate estimation analysis. Residues that comprise pockets of the PBR: A: 5, 7, 59, 63, 66, 99, 159, 163, 167, 171; B: 7, 9, 24, 25, 34, 45, 63, 66, 67, 70, 99; C: 9, 70, 73, 74, 97; D: 99, 113, 114, 155, 156, 159, 160; E: 97, 114, 133, 147, 152; F: 77, 80, 81, 84, 116, 123, 143, 416. Genbank accession numbers XX-XX.

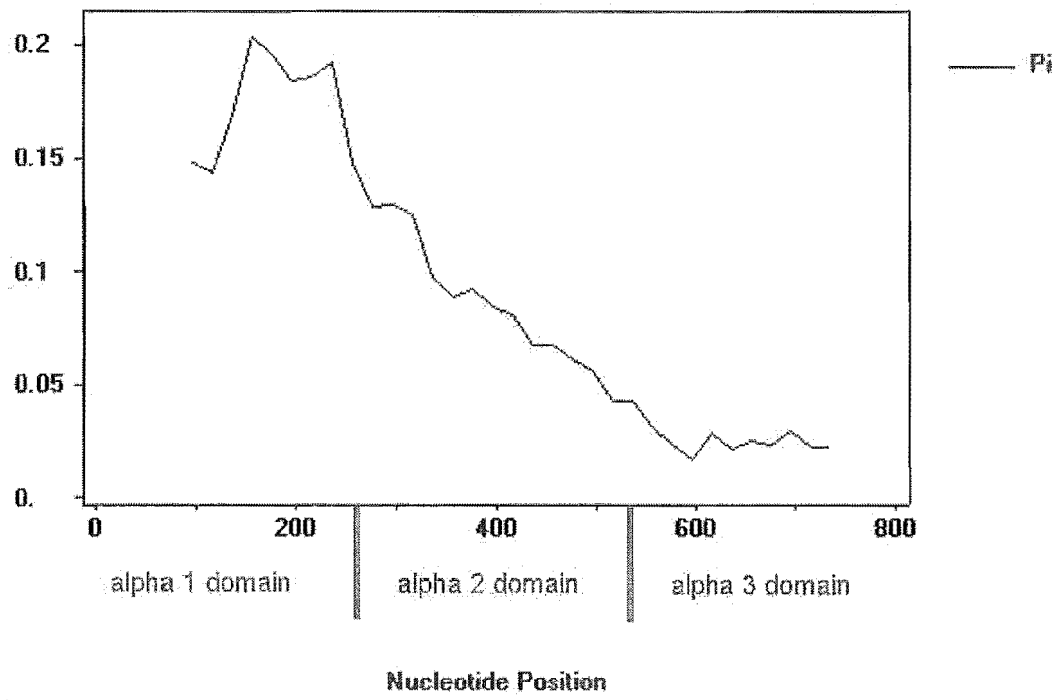


Figure 3.2. Sliding window analysis of nucleotide diversity in *Xenopus* MHC class Ia alleles.

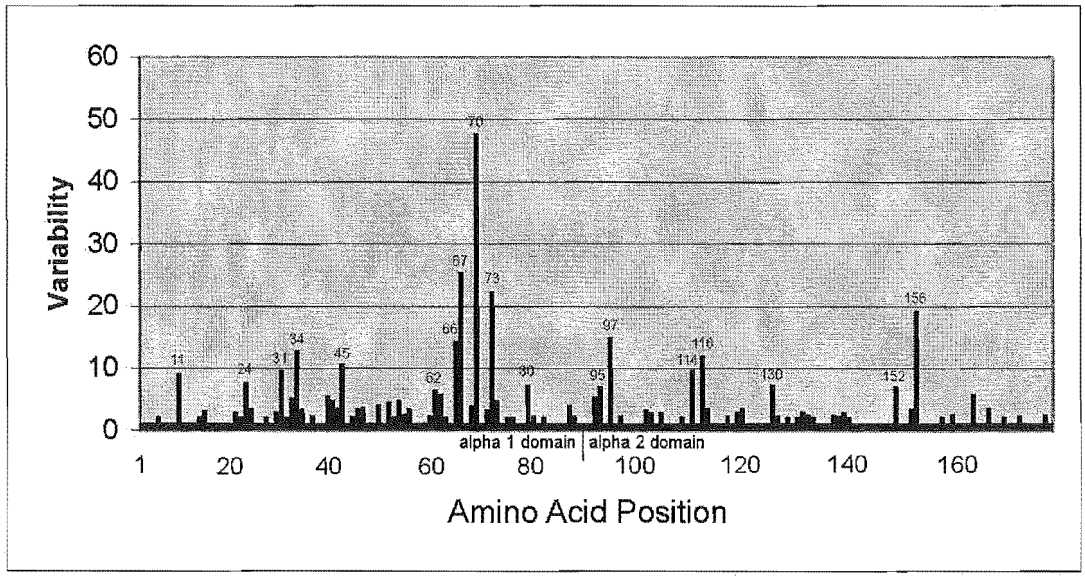


Figure 3.3. Wu-Kabat (1970) plot of amino acid variability in the α1 and α2 domains of the *X. laevis* class Ia MHC. Highly polymorphic sites have a Wu-Kabat score > 6 and are marked with their position corresponding to the HLA-A *0201 alignment.

divided in number between the β-sheets of the α1 and α2 domains, and all point into the cleft of the PBR. Eleven of the *X. laevis* highly polymorphic sites are in common with corresponding sites in the human MHC (Parham et al. 1988). However, 7 highly polymorphic sites in *X. laevis* (11, 31, 34, 73, 80, 130 and 152) are different from 6 sites (9, 65, 71, 74, 77 and163) that are highly polymorphic in humans but not *X. laevis*. Overall, patterns of amino acid variability correspond with results of nucleotide variation in that the α1 domain shows markedly higher levels of polymorphism that the α2 domain, especially in the α-helix regions of the PBR.

Certain sites of the α1 and α2 domains that comprise the PBR form 6 pockets or depressions within the binding cleft (Saper et al. 1991). Specificity of the peptide ligand is determined to varying degrees by each of the pockets, with pockets B and F having the largest impact. In pocket A, four of ten residues are completely conserved, and another three residues have very limited polymorphism. This pocket is polar in nature and tyrosine predominates at 4 sites. In contrast, pocket B is highly polymorphic. Five or more different residues are observed among alleles at five of eleven sites (34, 45, 66, 67, and 70) that comprise this pocket (Figure 3.1). Amino acid side chains at variable sites in this pocket have differing properties, accommodating a wide variety of peptide side chains. Pocket C is also polymorphic with over half of the sites (70, 73, 97) containing six or more amino acids among alleles. Residues of pocket D are mostly conserved and have polar side chains. This stands in contrast to the nature of this pocket in humans, where four of seven residues are polymorphic and the pocket is predominantly hydrophobic (Saper et al. 1991). The E pocket has two highly polymorphic sites one of which (114) contains predominantly polar residues and the other (97) contains mixture of polar and non-polar amino acids. In the F pocket two key positions (80 and 116) that determine peptide specificity (Zhang et al. 1998a) are variable, while other residues have limited polymorphism or are completely conserved. Polymorphisms at position 80 are largely conserved, while those at position 116 have side-chains of substantial steric difference.

SUBSTITUTION RATE ESTIMATION

From the frog MHC alleles in *Xenopus*04 comprising twenty-one sequences, six sequences were removed for substitution rate analysis because of recombination and one sequence was removed because of gaps in the 5' end of the α1 domain (Figure 3.1). After removing gaps, *Xenopus*99 was 921 bp long, or 307 codons; *Xenopus*04 was shorter and is

comprised of 260 codons because only the α1, α2, and α3 domain sequences were available. The relative fit of substitution models to *Xenopus*99 differed from that to *Xenopus*04. M8 was selected as the best approximating model for *Xenopus*04, while M3.1 was selected for *Xenopus*99 (Tables 3.1 and 3.2). Relative to the best approximating model, the fit of other models that allow for $d_n > d_s$ is much better than those that do not for both data sets. For both *Xenopus*99 and *Xenopus*04, model comparison shows that the fit of M7 is considerably better than that of M0, indicating that considerable variation exists in the ω ratios among codons in the sequence.

The parameter estimates of M8 and M3.1 in both data sets indicate that the majority of sites are moderately or highly conserved, but rate categories are present in which $\omega > 1$ (Tables 3.1 and 3.2). In the best approximating models, approximately 7 to 10% of sites fall into categories with $\omega > 1$. Other candidate models with unrestrained ω estimates also reveal a proportion of codon sites with $\omega > 1$. *Xenopus*99 showed similar patterns to *Xenopus*04, but the ω ratios of *Xenopus*99 were larger and the proportion of sites with elevated ω ratios was smaller. For models with codon sites $\omega > 1$, the codon sites estimated to be in this rate category are listed in Table 3.3. The sites listed are not evenly distributed throughout the length of the sequence, but are clustered in exons 2 and 3.

Model	D	lnL	δAIC	Parameter estimates
M8	37	-3535.814	Best	$p_0 = 0.926$, $p = 0.526$, $q = 0.991$ $p_1 = 0.074$, $\omega = 6.545$
M3	38	-3535.845	4.062	$p_0 = 0.566$, $p_1 = 0.362$, $p_2 = 0.072$ $\omega_0 = 0.104$, $\omega_1 = 0.747$, $\omega_2 = 6.699$
M3.1	40	-3535.322	7.016	$p_0 = 0.188$, $p_1 = 0.539$, $p_2 = 0.205$, $p_3 = 0.068$ $\omega_0 = 0.000$, $\omega_1 = 0.249$, $\omega_2 = 1.05$, $\omega_3 = 6.976$
M7	35	-3598.244	122.86	$p = 0.329$, $q = 0.527$, $\omega = 0.384$
M0	34	-3692.686	309.744	$\omega = 0.444$

Table 3.1. Likelihood values (lnL), number of free parameters in candidate models (D), differences in Akaike Information Criterion score (δAIC), and parameter estimates for *Xenopus*04

STRUCTURAL MODELING

The search for most appropriate template structures in the PDB showed that the F allele was best modelled by sequences from HLA-B locus, while the G allele was most similar

Model	D	lnL	δAIC	Parameter estimates
M3.1	34	-4054.934	best	$p_0 = 0.340$, $p_1 = 0.586$, $p_2 = 0.063$, $p_3 = 0.012$ $\omega_0 = 0.047$, $\omega_1 = 0.654$, $\omega_2 = 5.863$, $\omega_3 = 34.062$
M8	31	-4061.133	6.398	$p_0 = 0.949$, $p = 0.419$, $q = 0.466$ $p_1 = 0.051$, $\omega = 10.00$
M3	32	-4061.315	8.762	$p_0 = 0.397$, $p_1 = 0.548$, $p_2 = 0.053$ $\omega_0 = 0.069$, $\omega_1 = 0.754$, $\omega_2 = 9.677$
M7	29	-4128.993	138.118	$p = 0.367$, $q = 0.427$, $\omega = 0.463$
M0	28	-4208.673	295.496	$\omega = 0.502$

Table 3.2. Likelihood values (lnL), number of free parameters in candidate models (D), differences in Akaike Information Criterion score (δAIC), and parameter estimates for data of *Xenopus*99.

to nonclassical loci of the mouse (Qa-2). Structural features of the MHC amino acid backbone are generally conserved between lineages. One notable exception is the vertical portion of the α2 helix in allele G, which extends higher than in the F allele and forms a loop structure rather than a simple kink found in most MHC structures (see Appendix 2). As a result, the α2 helix of the G allele extends further than the peak of the α1 helix, whereas the two helices are similar in the F allele. The α1 helix appears to be similar in both lineages, as does the β sheet forming the floor of the PBR. Minor differences in loop regions outside the peptide binding regions are also evident. However, structural differences of loops in the backbone are not expected to appreciably alter the respective antigen repertoires of different lineages.

Side chains of class Ia sites exposed to residues of bound peptide show different levels of surface contact. There are seven residues of the F allele template that have extensive contact with the P2 side chain; three other amino acids have more limited contact (Table 3.4). Similar residues of the F and G allele templates bind with P2, indicating a largely conserved structure. Position 163 has a substantial surface area in contact with P2 in the F but not the G template. Residues at position 67, and to a lesser degree, position 45 in the MHC also have noticeably different levels of surface area in contact with the P2 side chain when compared across lineages. Many reported contact residues are highly polymorphic in both lineages indicating a variety of binding specificities among alleles. However, no consistent differences among lineages are found in polymorphic residues in contact with P2. While many sites in the α1 domain are

polymorphic and have elevated substitution rates, this structure lacks "constant but different" codon sites that may represent functional differences among lineages (Gribaldo et al. 2003).

MHC residues that make contact with PΩ differ more among lineages than do their P2 counterparts. In the F allele template, 14 residues bind with PΩ, while 11 residues make contact in the G allele template (Table 3.4). Many of the same amino acid positions make contact in both lineages, but several have different amounts of surface area exposed to bind PΩ. Most notably, MHC positions 77, 81, 116, and 143 have different levels of contact with the peptide. In the F allele template, positions 77, 116 and 143 have more surface area contact with PΩ, while the opposite is true for position 81. Most residues comprising PΩ contacts are conserved in *X. laevis* lineages, as they are known to bind the C terminal end of the peptide and are not intimately involved in side chain specificity (Zhang et al. 1998a). However, positions 95 and 116 of the F pocket are polymorphic (Figure 3.1), and these residues influence side chain specificity of the peptide (Zhang et al. 1998a).

Codon sites with ω > 1	
*Xenopus*04	*Xenopus*99
31	**24**
34	31
43	**34**
45	**43**
48	**45**
67	**62**
70	63
73	**66**
80	**67**
89	**70**
94	**73**
95	80
97	90
114	94
116	**97**
144	114
153	**153**
156	155
	156
	167
	170

Table 3.3. Positively selected codon sites ($P = 0.50$) estimated under the best approximating model for each data set (sites in boldface font $P = 0.05$). All selected sites except one are found in the PBR; numbering of sites follows the alignments shown in Figure 3.1.

DISCUSSION

POLYMORPHISM

Results of polymorphism in *X. laevis* MHC class Ia alleles reveal important and unexpected trends. The α1 and α2 domains might be expected to have similar levels and patterns of divergence and polymorphism because they are similar in and structure and function. However, our data demonstrate this is not the case in *X. laevis*. This trend has also been previously noted in other humans and other organisms (Kaufman et al. 1992; Parham et al. 1988), but the magnitude of difference between class Ia domains *X. laevis* is exceptional. Also, differences in relative levels of polymorphism among domains are not

the same among taxa. In humans, the helix of the α1 domain has more highly polymorphic sites than the helix of the α2 domain, and the reverse is true of the β strands

P2 (Glu)		P2 (Leu)		PΩ (Phe)		PΩ (Leu)	
F template		G template		F template		G template	
contact residue	surface area Å^2	contact residue	surface area Å^2	contact residue	surface area Å^2	contact residue	surface area Å^2
7	30.3	7	36.4	74	2.2	77	31.3
9	36.9	9	22.0	76	0.7	80	24.0
24	6.9	24	5.4	77	45.5	81	29.6
45	27.0	45	12.1	80	24.5	84	21.1
63	21.1	63	32.8	81	6.7	95	24.0
66	24.2	66	14.0	84	21.3	116	10.8
67	5.7	67	39.9	95	30.1	118	0.7
99	18.9	70	5.2	116	35.2	123	37.0
159	3.6	99	3.6	117	0.9	143	26.4
163	15.8	159	0.9	118	1.6	146	27.4
				123	35.4	147	18.8
				143	40.2		
				146	29.2		
				147	22.7		

Table 3.4. MHC residues and amount of surface area in contact with P2 and PΩ of the peptide antigen.

of the two domains. However, in *X. laevis*, the β strands of the two domains have equivalent numbers of highly polymorphic sites, and the α1 helix is more polymorphic than the α2 helix. Nevertheless, maintenance of differences in the relative levels of polymorphism among PBR domains from divergent taxa argues that the functions of these domains and selective pressures from various molecular interactions are conserved.

Class Ia proteins interact with a host of different molecules both inside and outside of the cell. Many of these interactions likely have a role in shaping the disparity in polymorphism among PBR domains. Most notably, the MHC is known to interact with the peptide, B_2 microglobulin, T-cell receptor (TCR) and CD8 co-receptor (Gao et al. 1997; Garboczi et al. 1996). The B_2 microglobulin and peptide both have similar levels of contact with the two domains of the PBR, and so may not have a large effect on differences in levels of polymorphism of the α1 and α2 domains. However, differences among domains may reflect divergent roles in binding peptides versus determining peptide specificity. The TCR forms contact with both the α1 and α2 domains, but the α2 domain has almost twice as many residues that bind with the TCR as the α1 domain. Also, the

CD8 receptor interacts strongly with three α2 sites, but has no interaction with the α1 domain. Therefore, these molecules may differentially influence the conservation of amino acid sites in the PBR domains. Further, the class Ia forms α2 domain-biased inter-molecular contacts with additional proteins in the ER.

Peptide loading complex proteins interact intimately with the class Ia glycoprotein (Pamer and Cresswell 1998). Early contact with the class I molecule is with the chaperones calnexin (CXN) and CRT, which bind in different ways and with different specificities (Harris et al. 1998). CRT binds in a specific manner primarily to areas of the α2 and α3 domains, but CXN binds in a non-specific manner. Likewise, TAP proteins, TPN and ERp57 seem to bind cooperatively to large areas of the α2 domain in a specific manner, but have nominal α1 domain contact (Yu et al. 1999). Interaction of the peptide loading complex proteins is important to class I operation and loss of these interactions can lead to severely reduced or loss of function; therefore amino acids associated with the peptide loading complex experience added selective pressure. The multiple association of the α2 domain with intracellular and extracellular proteins support the notion of differential conservation of the PBR domains by coevolution or interaction of functionally related molecules (Kaufman et al. 1992).

FUNCTIONAL DIFFERENCE OF LINEAGES

Structural features of models exhibit noticeable differences in steric properties of PΩ binding residues. In general, *X. laevis* alleles carry at position 116 an amino acid with either a large, aromatic side chain or a charged side chain. The residue at position 116 has a large impact on the peptide binding repertoire and a direct functional effect on disease and transplant pathology (Carrington and O'Brien 2003; Ferrara et al. 2001; Hulsmeyer et al. 2002; Kubo et al. 1998; Zhang et al. 1998a). In *X. laevis*, steric and charge differences among alleles at this position likely has a large impact on peptide binding repertoires. I propose that F pocket differences represent a fundamental division between *X. laevis* MHC class Ia alleles. However, rigorous experimental examination of functional differences between lineages of *X. laevis* is not yet available and inference is drawn based on observational data. Nevertheless, the steric differences likely cause a divergence in the peptide repertoires of alleles that is significant enough to warrant the mutual maintenance of divergent allelic lineages in evolutionary time.

In the F pocket, differences among alleles in residues and amount of surface area that are in contact with PΩ are apparent. A comparison of the total surface area in contact with P7 and PΩ reveals differing levels of association among alleles. The F allele template has 296 $Å^2$ contact area with PΩ, whereas the G template has 251 $Å^2$ surface area contact with that position. In contrast, the F and G allele templates have 135 $Å^2$ and 161 $Å^2$ respectively of contact with P7. In lineage A, P7 may play a more prominent role in binding and determining peptide specificity. These trends underscore the relative influence that these peptide residues have on MHC restriction and may represent a fundamental difference in how various lineages bind peptides. These findings provide evidence supporting the hypothesis that the lineages function to present different sets of peptides and that they bind peptides in different ways.

CO-EVOLUTION OF MHC REGION GENES

A basis for co-evolution of class I region genes can be found in the close association of MHC processing proteins with class I proteins in the ER, and in variation at position 116 (Beißbarth et al. 2000; Hildebrand et al. 2002; Turnquist et al. 2000). Differences at position 116 in humans have a strong effect on associations and efficiency of class I protein interactions with processing proteins (Neisig et al. 1996; Williams et al. 2002). In *Xenopus*, the manner of optimal peptide loading may differ for the two lineages. Use of cellular machinery from one lineage to load peptides onto an MHC allele of the other lineage may then result in markedly diminished immuno-surveillance and CTL response (Hildebrand et al. 2002). Thus variation at position 116 in *Xenopus* MHC lineages may contribute co-evolutionary tendencies of linked MHC processing genes due to their close interaction and optimization of antigen repertoire. In addition, pressure to transport and load different sets of peptides may contribute to maintenance of allelic lineages and co-evolution of MHC processing genes (Joly et al. 1998).

Other species have shown co-evolutionary patterns between class I lineages and physically linked MHC processing genes that are dependent on F pocket variation (Joly et al. 1998; Kaufman 1999). In both the rat and chicken, co-adapted MHC gene complexes differ in C terminus peptide specificity as is the case with proposed differences in *Xenopus* lineages. Co-evolutionary tendencies only are possible where close linkage and a lack of recombination have allowed loci to co-segregate (Kaufman 1999). In *X. laevis*, evidence indicates that the class I region is partially comprised of closely linked class I loci and

class I processing genes such as *tap1, tap2* and *lmp7* (Nonaka et al. 1997a; Ohta et al. 1999; Ohta et al. 2003). Although recombination rates between loci have not been rigorously investigated, the segregation patterns predict that the crossover rate will be low. I note that the genetic exchange detected within the MHC class I locus detected here represents small-scale genetic exchange likely due to gene conversion and is not expected to alter inter-locus linkage patterns (Andolfatto and Nordborg 1998).

NATURAL SELECTION OF XENOPUS MHC

Convincing evidence for molecular evolution by natural selection comes from the comparison of d_n and d_s estimates (Sharp 1997; Yang and Bielawski 2000). I show for the first time that for 7% -10% of sites in these data, the nonsynonymous substitution rate exceeds the synonymous substitution rate. These patterns of substitutions are most easily interpreted as the result of natural selection and provide evidence that natural selection at the MHC class Ia gene of *X. laevis* and is operating to increase genetic diversity and maintain allelic lineages. The location of inferred sites under natural selection also favours the hypothesis of natural selection. With few exceptions, the sites under selection in these data are found in the exons coding for the PBR. The lack of complete correspondence in selected sites and essential α and β structures may be due to slight differences between taxa in the amino acids that make up these features, or from spurious identification of selected sites owing to the difficulty of identifying selected sites (Anisimova et al. 2002).

Workers have experienced difficulty demonstrating natural selection when considering some other MHC gene data sets. For instance, both the salmonid and other *Xenopus* species have very high levels of diversity and lower d_n/d_s values (Sammut et al. 2002; Shum et al. 2001). In both cases, the mode and level of polymorphism were thought to contribute to the low d_n/d_s values. These instances involve polyploid species, and polypolidization often leads to higher rates of evolution and extensive and rapid genome change (Soltis and Soltis 1995), masking evidence of natural selection. I estimated substitution rates in salmonid fishes and found that in both cases evidence exists for the operation of natural selection on a number of codon sites (brown trout: $p_2 = 0.104$ $\omega_2 = 3.94$; rainbow trout: $p_2 = 0.072$, $\omega_2 = 5.65$). Thus as in *X. laevis*, previously obtained conservative estimates of d_n/d_s in other species may represent an artefact of rate estimators undercounting the actual numbers of substitutions.

Comparison of the substitution rate analysis between the two frog data sets reveals some differences. Although conclusions regarding selection and elevation of substitution rates are consistent between data sets, one data set produces higher values than the other. These differences may be due to recombination in one data set but not the other. The overall effect of recombination is to increase the number of false positives, artificially raising the rate estimates (Anisimova et al. 2003; Shriner et al. 2003). Generally, each data set is expected to have attributes that are particular to those data. However, I wish to draw inferences regarding the overall underlying process of MHC class Ia evolution in *X. laevis* rather than about a particular data set. This is achieved through careful model selection using the AIC, so that models do not over-parameterize the information in the data (Burnham and Anderson 2002). In doing so, these conclusions offer some generality with respect to the relationship of natural selection, functional differences and co-evolution in the MHC of *X. laevis*.

ACKNOWLEDGEMENTS:

Thanks to Martin Flajnik, Neil Gemmell, and anonymous peer reviewers for suggestions on earlier versions of this manuscript. David Bos is supported by the Marsden Fund of New Zealand and a PhD scholarship from the University of Canterbury.

REFERENCES

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control AC 19:716-723

Andolfatto P, Nordborg M (1998) The effect of gene conversion on intralocus associations. Genetics 148:1397-1399

Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of bayes prediction of amino acid sites under positive selection. Molecular Biology and Evolution 19:950-958

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229-1236

Apanius V, Penn DJ, Slev P, Ruff LR, Potts WK (1997) The nature of selection on the Major Histocompatibility Complex. CRC Critical Reviews in Immunology 17:179-224

Beiβbarth T, Sun J, Kavathas PB, Ortmann B (2000) Increased efficiency of folding and peptide loading of mutant MHC class I molecules. European Journal of Immunology 30:1203-1213

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Research 28:235-242

Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference: a practical information-theoretic approach. Springer-Verlag, New York

Carrington M, O'Brien SJ (2003) The influence of *HLA* genotype on AIDS. Annual Review of Medicine 54:535-551

Crandall KA, Kelsey CR, Imamichi H, Lane HC, Salzman NP (1999) Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. Molecular Biology and Evolution 16:372-382

Ferrara GB, Bacigalupo A, Lamparelli T, Lanino E, Delfino L, Morabito A, Parodi AM, Pera C, Pozzi S, Sormani MP, Bruzzi P, Bordo D, Bolognesi M, Bandini G, Bontadini A, Barbanti M, Frumento G (2001) Bone marrow transplantation from unrelated donors: the impact of mismatches with substitutions at position 116 of the human leukocyte antigen class I heavy chain. Blood 98:3150-3155

Flajnik MF, Kaufman J, Riegert P, Du Pasquier L (1984) Identification of class I major histocompatibility complex encoded molecules in the Amphibian *Xenopus*. Immunogenetics 20:134-143

Flajnik MF, Ohta Y, Greenberg AS, Salter-Cid L, Carrizosa A, Du Pasquier L, Kasahara M (1999) Two ancient allelic lineages at the single classical class I locus in the *Xenopus* MHC. Journal of Immunology 163:3826-3833

Gao GF, Tormo J, Gerth UC, Wyer JR, McMichael AJ, Stuart DI, Bell JI, Jones EY, Jakobsen BK (1997) Crystal structure of the complex between human CD8αα and HLA-A2. Nature 387:630-634

Garboczi DN, Ghosh P, Utz U, Fan QR, Biddison WE, Wiley DC (1996) Structure of the complex between human T-cell receptor, viral peptide and HLA-A2. Nature 384:134-141

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11:725-736

Grandea III AG, Van Kaer L (2001) Tapasin: an ER chaperone that controls MHC class I assembly with peptide. Trends in Immunology 22:194-199

Gribaldo S, Casane D, Lopez P, Philippe H (2003) Functional Divergence Prediction from Evolutionary Analysis: A Case Study of Vertebrate Hemoglobin. Molecular Biology and Evolution 20:1754-1759

Guex N, Diemand A, Peitsch MC (1999) Protein modeling for all. Trends in Biochemical Sciences 29:364-367

Guex N, Peitsch MC (1997) SWISS-MODEL and the Swiss-Pdbviewer: an environment for comparative protein modeling. Electrophoresis 18:2714-2723

Harris MR, Yu YYL, Kindle CS, Hansen TH, Solheim JC (1998) Calreticulin and Calnexin Interact with Different Protein and Glycan Determinants During the Assembly of MHC Class I. J Immunol 160:5404-5409

Hashimoto K, Okamura K, Yamaguchi H, Ototake M, Nakanishi T, Kurosawa Y (1999) Conservation and diversification of MHC class I and its related molecules in vertebrates. Immunological reviews 167:81-100

Hildebrand WH, Turnquist HR, Prilliman KR, Hickman HD, Schenk EL, McIlhaney MM, Solheim JC (2002) HLA class I polymorphism has a dual impact on ligand binding and chaperone interaction. Human Immunology 63:248-255

Hill AVS (1999) Defence by diversity. Nature 398:668-669

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170

Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex loci of vertebrates. Annual Review of Genetics 32:415-435

Hulsmeyer M, Hillig RC, Volz A, Ruhl M, Schroder W, Saenger W, Ziegler A, Uchanska-Ziegler B (2002) HLA-B27 Subtypes Differentially Associated with Disease Exhibit Subtle Structural Alterations. Journal of Biological Chemistry 277:47844-47853

Ina Y (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. Journal of Molecular Evolution 40:190-226

Joly E, Le Rolle AF, Gonzalez AL, Mehling B, Stevens J, Coadwell WJ, Hunig T, Howard JC, Butcher GW (1998) Co-evolution of rat TAP transporters and MHC class I RT1-A molecules. Current Biology 8:169-172

Kaufman J (1999) Co-evolving genes in MHC haplotypes: the "rule" for nonmammalian vertebrates? Immunogenetics 50:228-236

Kaufman J, Anderson R, Avila D, Engberg J, Lambris J, Salomonsen J, Welinder K, Skjodt K (1992) Different features of the MHC class I heterodimer have evolved at different rates. Journal of Immunology 142:1532-1546

Kaufman J, Salomonsen J, Flajnik MF (1994) Evolutionary conservation of MHC class I and class II molecules--different yet the same. Seminars in Immunology 6:411-424

Kubo H, Ikeda-Moore Y, Kikuchi A, Miwa K, Nokihara K, Schonbach C, Takiguchi M (1998) Residue 116 determines the C-terminal anchor residue of HLA-B*3501 and -B*5101 binding peptides but does not explain the general affinity difference. Immunogenetics 47:256-263

Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. Bioinformatics 16:562-563

Muse SV (1996) Estimating synonymous and nonsynonymous substitution rates. Molecular Biology and Evolution 13:105-114

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Molecular Biology and Evolution 11:715-724

Namikawa C, Salter-Cid L, Flajnik MF, Kato Y, Nonaka M, Sasaki M (1995) Isolation of Xenopus *LMP-7* homologues: striking allelic diversity and linkage to MHC. Journal of Immunology 155:1964-1971

Neisig A, Wubbolts R, Zang X, Melief C, Neefjes J (1996) Allele-specific differences in the interaction of MHC class I molecules with transporters associated with antigen processing. Journal of Immunology 156:3196-3206

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and application to the HIV-1 envelope gene. Genetics 148:929-936

Nonaka M, Namikawa C, Kato Y, Sasaki M, Salter-Cid L, Flajnik MF (1997a) Major histocompatibility complex gene mapping in the amphibian *Xenopus* implies a primordial organization. Proceedings of the National Academy of Sciences, USA 94:5789-5791

Nonaka M, Namikawa-Yamada C, Sasaki M, Salter-Cid L, Flajnik MF (1997b) Evolution of proteasome subunits δ and LMP2. Journal of Immunology 159:734-740

Ohta Y, McKinney EC, Criscitiello MF, Flajnik MF (2002) Proteosome, Transporter associated with antigen processing, and class I genes in the Nurse Shark *Ginglymostoma cirratum:* evidence for a stable class I region and MHC haplotype lineages. Journal of Immunology 168:771-781

Ohta Y, Powis SJ, Coadwell WJ, Haliniewski DE, Liu Y, Li H, Flajnik MF (1999) Identification and mapping of *Xenopus* TAP2 genes. Immunogenetics 49:171-182

Ohta Y, Powis SJ, Lohr RL, Nonaka M, Du Pasquier L, Flajnik MF (2003) Two highly divergent ancient allelic lineages of the transporter associated with antigen processing (TAP) gene in *Xenopus*: further evidence for co-evolution among MHC class I region genes. European Journal of Immunology 33:3017-3027

Pamer E, Cresswell P (1998) Mechanisms of MHC class I-restricted antigen processing. Annual Review of Immunology 16:323-358

Parham P, Lomen CE, Lawlor DA, Ways JP, Holmes N, Coppin HL, Salter RD, Wan AM, Ennis PD (1988) Nature of polymorphism in HLA-A, -B, -C molecules. Proceedings of the National Academy of Sciences USA 85:4005-4009

Parham P, Ohta T (1996) Population biology of antigen presentation by MHC class I molecules. Science 272:67-74

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818

Potts WK, Slev P (1995) Pathogen-based models favoring MHC genetic diverstiy. Immunological reviews 143:181-196

Rozas J, Rozas R (1999) DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics 15:174-175

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425

Sammut B, Marcuz A, Du Pasquier L (2002) The fate of duplicated major histocompatibility complex class Ia genes in a dodecaploidd amphibian, *Xenopus ruwenzoriensis*. European Journal of Immunology 32:1593-1604

Saper MA, Bjorkman PJ, Wiley DC (1991) Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 A resolution. Journal of Molecular Biology 219:277-319

Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Research 31:3381-3385

Sharp PM (1997) In search of molecular Darwinism. Nature 385:111-112

Shriner D, Nickle DC, Jensen MA, Mullins JI (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. Genetical Research Cambridge 81:115-121

Shum BP, Guethlein LA, Flodin LR, Adkinson MA, Hedrick RP, Nehring RB, Stet RJM, Secombes C, Parham P (2001) Modes of Salmon MHC class I and II evolution differ from the primate paradigm. Journal of Immunology 166:3297-3308

Sobolev V, Sorokine A, Prilusky J, Abola E, Edelman M (1999) Automated analysis of interatomic contacts in protiens. Bioinformatics 15:327-332

Soltis DE, Soltis PS (1995) The dynamic nature of polyploid genomes. Proceedings of the National Academy of Sciences, USA 92:8089-8091

Suzuki Y, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. Molecular Biology and Evolution 16:1315-1328

Swofford DL (1998) PAUP* Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.0. Sinauer, Sunderland, MA

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673-4680

Turnquist HR, Thomas HJ, Prilliman KR, Lutz CT, Hildebrand WH, Solheim JC (2000) HLA-B polymorphism affects interactions with multiple endoplasmic reticulum proteins. European Journal of Immunology 30:3021-3028

Williams AP, Au Peh C, Purcell AW, McClusky J, Elliot T (2002) Optimization of the MHC class I peptide cargo is dependent on tapasin. Immunity 16:509-520

Wu TT, Kabat EA (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and meyloma light chains and their implications for antibody complimentarity. Journal of Experimental Medicine 132:211-250

Yang Z (2000) Phylogenetic Analysis by Maximum Likelihood (PAML). University College London, London

Yang Z (2001) Adaptive Molecular Evolution. In: Balding DJ, Cannings C, Bishop M (eds) Handbook of Statistical Genetics. John Wiley and Sons, New York, p 327-350

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution 15:496-502

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular Biology and Evolution 17:32-43

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution 19:908-917

Yang Z, Nielsen R, Goldman N, Pedersen A-MK (2000) Codon substitution models for heterogeneous selection pressure and amino acid sites. Genetics 155:431-449

Yu YYL, Turnquist HR, Myers NB, Balendiran GK, Hansen TH, Solheim JC (1999) An extensive region of an MHC class I $\alpha 2$ domain loop influences interaction with assembly complex. Journal of Immunology 163:4427-4433

Zhang C, Anderson A, DeLisi C (1998) Structural principles that govern the peptide-binding motifs of class I MHC molecules. Journal of Molecular Biology 281:929-947

CHAPTER IV

MODE OF MHC CLASS IA EVOLUTION IN *XENOPUS LAEVIS*

ABSTRACT

The mode of MHC evolution involves duplications, deletions and independent divergence of loci during episodes punctuated by natural selection. Major differences in the mode of MHC evolution among taxa have been attributed to genomic linkage patterns of class I and class II MHC genes. Here, I characterize the mode of evolution in the classical class Ia MHC gene of *Xenopus laevis*, and test whether or not the independence of class I and class II genes is necessary for particular patterns of class I gene evolution reported for salmonid fishes. In *X. laevis*, genetic exchange is relatively frequent and occurs in intron II, reshuffling allelic forms of exons 2 and 3. This finding is similar to results reported for salmonid fishes but differs from the pattern common to the mammalian paradigm of MHC evolution. Evolutionary relationships among class I alleles show an intermingling of alleles from different *Xenopus* species, rather than a species-specific clustering. Results indicate that the mode of evolution is similar to that found in salmonid fishes and is different than the mode of evolution seen in primates. Known linkage of MHC region genes in *X. laevis* suggest that the mode of evolution common to salmonid fishes and *X. laevis* is not due primarily to nonlinkage of MHC class I and class II regions.

INTRODUCTION

MHC class I and class II genes are found in all gnathostomes and encode structurally similar proteins that present antigenic peptides to T lymphocytes. (Abbas et al. 2000). Class II proteins are expressed mainly on specialized antigen presenting cells and function to bind peptides derived from extracellular pathogens. Class I proteins are expressed in almost all cells and are involved in monitoring the internal environment of the cell for foreign, mutated or misfolded proteins. Class I genes are subdivided into classical (class Ia) and nonclassical (class Ib) loci based on differences in polymorphism, structure, function and expression pattern (Parham 1994). Aside from their important role in the immune system, the MHC genes are of particular interest because of their unusual evolutionary genetic features.

Phylogenetic and population genetic comparison of class I and class II alleles from various taxa has revealed several extraordinary characteristics, including a high level of amino acid polymorphism and allelic diversity at the population level (Parham and Ohta 1996). Some MHC allelic lineages also exhibit unusual longevity which predates the formation of species (Figueroa et al. 1988; Lawlor et al. 1988). As a result, "trans-species polymorphisms" exist, whereby some MHC alleles are more closely related to alleles from other species rather than from within the same species. Variation in linkage patterns, order of gene loci, and the number of gene family members resulting from tandem duplications has also been observed (Trowsdale 1995). Interallelic recombination in the mammalian MHC has also been detected as a mechanism generating allelic diversity (Jeffreys et al. 2001; Martinsohn et al. 1999). Finally, many of the features of the MHC have been attributed to the forces of balancing selection acting at the molecular level (Hughes and Yeager 1998).

The mode of MHC evolution has been most thoroughly investigated among primate taxa. In the classical paradigm, class I and class II families of genes evolve differently from each other, each with its own rates of locus duplication, divergence and allelic longevity. Class I loci have a higher rate of duplication and replacement than class II gene loci, and many functional class I loci have arisen after speciation events (e.g. Cadavid and Watkins 1997). In primates, certain class II allelic lineages predate prosimian divergence from the proto-human lineage (Bontrop et al. 1999). In contrast, class I lineages are younger and extend only as far back as the emergence of great apes

(Vogel et al. 1999). Genetic exchange in humans is characterized by the intra-locus exchange of small mini-cassettes of nucleotides that can occur throughout the length of the gene and has no single prominent breakpoint (Hughes et al. 1993; Jakobsen et al. 1998). Until recently, it was not known whether these well established patterns of MHC evolution also were found among other species.

Despite many elements of conserved structure and function, non-mammalian taxa have different patterns of MHC evolution compared to primates. For instance, class II gene evolution in birds is different than mammals in that those loci have a relatively recent origin (Edwards et al. 1995; Hess and Edwards 2002). Also, class II allelic lineages in salmonid fishes cluster in a species-specific manner, while class Ia lineages share trans-species polymorphism among divergent taxa (Shum et al. 2001). Salmonid fishes also have much higher levels of intralocus class Ia polymorphism than the most polymorphic locus in humans. In general, recombination plays a more prominent role in teleost class Ia evolution as compared to mammals (Shum et al. 2001). Intragenic recombination in salmonids typically involves entire exons and a prominent breakpoint for genetic exchange is easily identifiable. In fishes, this mode of evolution has been attributed to the unique linkage patterns of class I and class II genes in bony fishes, which are found on two separate chromosomes (Bingulac-Popovic et al. 1997).

Xenopus laevis, the African clawed frog, is the first poikilothermic vertebrate from which class I proteins were isolated (Flajnik et al. 1984). In this species, there is a single MHC class Ia locus with diploid inheritance patterns (Shum et al. 1993). This locus has high levels of polymorphism when compared to mammals, but is similar to the variation found in salmonid fishes. Another unusual aspect of *Xenopus* class Ia is the existence of two ancient allelic lineages (Flajnik et al. 1999). These lineages are very distinct, as alleles belonging to different lineages are as divergent as MHC alleles from mouse and human. Linkage patterns of class Ia and class II genes in *X. laevis* indicate a single MHC genomic region like many vertebrates, but unlike bony fishes (Nonaka et al. 1997a).

Xenopus laevis provides an opportunity to investigate different modes of evolution in various taxa by exploiting similarities and differences in various patterns of MHC genetics. Linkage of certain genes within the class I region of the MHC are alike among fishes and frogs. However, the linkage pattern of the entire MHC is different in that *X. laevis* class I and class II regions are linked whereas they are not in teleost fishes. The pattern of recombination and extent of trans-species polymorphism among MHC class Ia

alleles of *Xenopus* species has not yet been examined. To further investigate the mode of MHC evolution in this species, I examine the effects of recombination on the creation and maintenance of allelic diversity in *X. laevis*. I also reconstruct phylogenetic relationships among class Ia alleles of various *Xenopus* species to investigate trans-species patterns of polymorphism. If the patterns of class Ia polymorphism seen in salmonid fishes is primarily due to the non-linkage of class I and class II genes, we would expect to observe the classical pattern of MHC evolution in *X. laevis* because those genes are linked. However, if MHC evolution in *X. laevis* is similar that of fishes then we can infer that non-linkage of class Ia and class II genes is not vital to the mode of MHC evolution seen in salmonids.

MATERIALS AND METHODS

DATA COLLECTION

I extracted total RNA from *X. laevis* blood samples using the TRIzol protocol following manufacturer's recommendations (Invitrogen). One μL total RNA was added to a Superscript One-Step RT_PCR kit (Invitrogen) and first strand synthesis was performed at 55° C for 25 min. Immediately after first strand synthesis, PCR was employed on cDNA with primers designed to amplify exons 1-3 of the MHC class Ia gene (forward primer: 5'-GTCACTCCCTGCGYTAYTAT-3'; reverse primer: 5'-TTTCTCCTTCAGGCTGCTGT-3'). Primers were designed using the Primer3 website (http://www-genome-wi.mit.edu/cgi-bin/primer/primer3_www.cgi) from known *X. laevis* sequences (Flajnik et al. 1999), and the PCR protocol was optimized to minimize the occurrence of *in vitro* recombination (Judo et al. 1998). I cloned PCR products into the pCR 4 TOPO TA plasmid following manufacturer's recommendations (Invitrogen), and recombinant DNA was transformed into TOP-10 *Escherichia coli* cells. *E. coli* cells were plated onto LB agar and grown overnight at 37° C after which 6-10 individual colonies were picked and grown in LB broth at 37° C for 16 h.

Five mL of LB broth/cell matrix was removed and plasmid DNA was extracted using alkaline lysis mini-preps (Sambrook et al. 1989). I sequenced the MHC insert in both directions using BigDye v3.1 chemistry and an ABI 3730 automated sequencer. ABI trace files were edited using Bioedit (Hall 1999) and sequences were aligned using Clustal W (Thompson et al. 1994). Eleven new sequences were isolated from eleven *X. laevis* chromosomes; these sequences were independently verified from 2-6 separate colonies.

Additional sequences were obtained but were not recovered multiple times and were excluded from the following analyses. New sequences were added to other known class Ia sequences of exons 2, 3 and 4 from frogs (Genbank accession numbers, *X. tropicalis:* AY204558, AY204559; *X. ruwenzoriensis:* AF497525 - AF497528; and *X. laevis*, *Rana pipiens*, and a laboratory-bred interspecies hybrid of *X. laevis-X. gilli:* AF185579-AF185588).

STATISTICAL ANALYSIS

I used various statistical methods to investigate evolutionary relationships and the effects of intragenic recombination. Maxchi was used to detect the occurrence of recombination events within *X. laevis* samples because it performs well in simulations and had a low false error rate (Posada and Crandall 2001a). The program RDP (Martin and Rybicki 2000) characterized intragenic recombination by identifying breakpoints and alleles created by recombination. The *P* value of significant differences used to infer a recombination was set at 0.000005 with a window size of 10 nucleotides to minimize the false positive error rate (Martin and Rybicki 2000). RDP allows the inference of a recombination break point, and that information was used during phylogenetic analyses.

Evolutionary relationships were reconstructed among known *Xenopus* MHC class Ia sequences. Prior to phylogenetic analysis, I tested for loss of information in these data due to saturation using the index of substitution saturation (Xia et al. 2003). The best approximating model of nucleotide evolution for these data was determined using Akaike's information criterion (AIC) (Akaike 1974). Maximum likelihood (ML) scores of candidate models were calculated using PAUP* 4.0 (Swofford 1998) and AIC scores computed in Modeltest (Posada and Crandall 1998). Employing the best approximating model, genetic distance and phylogenetic relationships were estimated using ML optimization.

A traditional bifurcating phylogenetic tree may not accurately represent evolutionary relationships among a population sample of DNA sequences because of genetic exchange (Posada and Crandall 2001b). Therefore, I reconstructed separate trees by partitioning these data into congruent segments with shared evolutionary history on either side of a putative recombination break point. Maximum likelihood (Felsenstein 1981) and Neighbor Joining (NJ Saitou and Nei 1987) reconstructions were performed to test hypotheses regarding lineage assortment among species. Maximum likelihood scores

were compared to a range of *a priori* topologies corresponding to different levels of trans-species lineage sharing among taxa. Tree comparisons were performed using Shimodaira-Hasegawa tests (Shimodaira 2002; Shimodaira and Hasegawa 1999) implemented in the program package Consel (Shimodaira and Hasegawa 2001). One thousand bootstrap pseudoreplications were used to estimate support for nodes in the NJ tree, and > 50% bootstrap support in resulting topologies are shown.

RESULTS

DATA

The critical value of the index of substitution saturation ($I_{SS.C}$) represents the index of substitution saturation (I_{SS}) value beyond which data fail to recover a true phylogenetic tree. The I_{SS} values for the α1 and α2/α3 domain partitions are 0.261 and 0.228. The $I_{SS.C}$ values are 0.682 and 0.715 respectively, and these are significantly larger than the respective I_{SS} scores ($P < 0.000$). Overall, the total data consist of 27 sequences and have 397 polymorphic sites out of 781 total nucleotides; on average, alleles differ by 99.30 nucleotides.

RECOMBINATION

Intragenic recombination plays a prominent role in *X. laevis* MHC evolution. Although substitutions are a major factor in the evolution of MHC, recombination is responsible for the creation of a number of new alleles. Overall, the number of alleles created through recombination is at least 6 out of 19, over 30% of alleles in this data set (Table 4.1). The parameters of RDP were set conservatively to avoid false positive identification of recombination events, so this number represents a minimum number of recombination alleles. The recombination break point also is shared among alleles, indicating that recombination is not free, but commonly occurs in

Sequence	breakpoint	parent sequences
Xela R	230	Xela 30.7 / unknown
Xela 14	260	lg a/c2 / Xela 39
lg b/d2	258	Xela 14 / unknown
Xela 30.7	250	lg a/c1 / Xela 14
Xela 39	256	Xela F / Xela 14
Xela 44	250	lg a/c1 / Xela 14

Table 4.1. *X. laevis* recombinant sequences. The sequences on the left are created through a genetic exchange event whose crossover event occurred at the breakpoint indicated. The two alleles inferred to combine to form the recombinant sequence are listed on the right, if known.

intron II. The size of the DNA fragment involved in the exchanges typically encompasses the entire α1 domain coding exon. This type of recombination leads to intralocus allelic "exon shuffling" that creates new arrangements of existing variation in the peptide binding region (PBR). This pattern is very different from that seen in humans, where genetic exchange involves much smaller fragments.

Some other trends in the pattern of recombination in *X. laevis* are noteworthy. For instance, some alleles are involved in genetic exchange more often than others. The occurrence of genetic exchange often involves allele 14 in these data. This allele is involved in creating four new alleles; lg a/c1 is a parent sequence for two additional recombinant alleles. A bias in alleles involved in recombination has also been detected in salmonid fishes (Shum et al. 2001). Inspection of recombinants reveals that two recombinant alleles (alleles 30.7 and 44) have identical α1 domain sequences, but the α2 and α3 domains of these two alleles are different. Finally, the formation of recombinant alleles is not restricted to closely related alleles, as two highly divergent alleles can recombine (e.g. alleles F and 14). The apparent ongoing genetic exchange results in a high level of recombination that is likely to affect the evolutionary relationships among alleles and different domains of alleles.

EVOLUTIONARY RELATIONSHIPS

The evolutionary relationships among MHC class Ia alleles in *Xenopus* species were determined using ML optimization and by NJ topology reconstruction. Tree reconstruction was done separately on two segments of the sequence, partitioning the α1 domain as one segment and the α2 and α3 domains together as the other segment. This partition was chosen to maximize detection of different evolutionary histories due to genetic exchange, and provides a means for confirming the presence of recombination in this data set. In the phylogenetic trees of the α1 and α2/α3 domains, recombinant alleles identified with the program RDP were found to be in different clades. These alleles moved across nodes with > 50% bootstrap support to associate with different sets of other alleles in each tree reconstruction. The translocation of alleles to different parts of the tree topology is consistent with patterns of recombination detected with RDP.

The best approximating model selected for α1 domains differs from that chosen for the α2/α3 sequences (Table 4.2). The model favoured to describe α1 domain sequence evolution is TIM+Γ (Posada and Crandall 2001c); for the α2/α3 sequence fragment the

K81uf+Γ (Kimura 1981) model was preferred. Compared to the α1 domain, the gamma shape parameter was smaller in the α2/α3 domain, indicating more rate variation in this part of the sequence. Specific nucleotide frequencies differed among partitions with adenine by far the most common in the α1 domain but guanine the most prevalent in α2/α3 domain. Relative rates in the substitution matrix are higher in the α1 domain, with the C↔T substitution rate more than twice the rate found in the α2/α3 sequence. Moreover, transversions are also found at a substantially higher rate in the α1 domain. Differences between sequence partitions are not limited to models of evolution, but are also found in tree topology, lending further endorsement to the data partition used here.

MHC domain	**α1**	**α2/α3**
Selected model	**TIM+Γ**	**K81uf+Γ**
Base frequency		
A	0.328	0.276
T	0.201	0.215
G	0.253	0.295
C	0.218	0.214
Substitution rates		
A↔C	1.00	1.00
A↔G	2.43	1.89
A↔T	1.41	0.75
C↔G	1.41	0.75
C↔T	3.89	1.89
G↔T	1.00	1.00
Rate variation		
α parameter	0.576	0.466

Table 4.2. Best approximating model parameters for data partitions. Different optimal models were selected for data partitions that separate the α1 and α2/α3 domains of the MHC molecule.

The topology showing relationships among α1 domain sequences shows mixing of alleles from different species (Figure 4.1). Pairs of alleles from a species form well supported groups in some cases, but both *X. tropicalis* and *X. ruwenzoriensis* α1 domains are intermingled together with *X. laevis* sequences. One group of *X. laevis* alleles is closely related and forms a tight cluster that has 100% bootstrap support; this group includes four recombinant sequences. In other parts of the topology, some terminal branches are long, as would be expected from sequences subject to balancing selection. Most well supported clades are near the tips of the tree and are comprised of only a few sequences; branches in the more basal parts of the tree are typically shorter than many terminal branches.

The evolutionary relationships reconstructed for α2/α3 domain sequences were different in some ways to the α1 domain sequence topology (Figure 4.2). For instance, some alleles segregated by species rather than being intermingled. In this tree *X.*

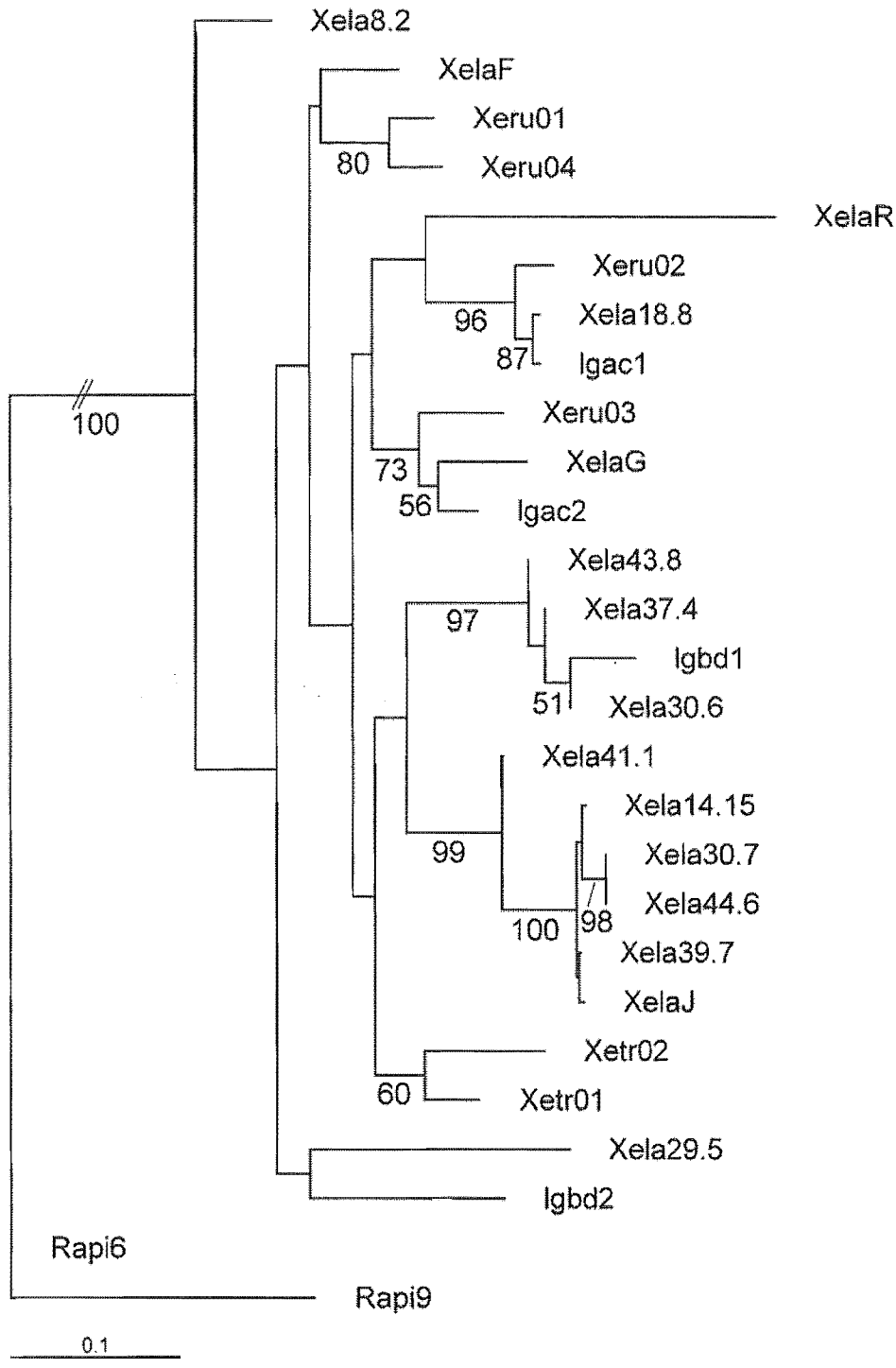


Figure 4.1. Evolutionary relationships of *Xenopus* class Ia sequences using α1 domain sequences. Numbers indicate bootstrap support for nodes. All branches shown to the scale on the bottom left of figure except the branch leading to the out group which was shortened for graphical clarity of the remaining branches of the tree. *Rapi R. pipiens; Xela X. laevis; Xetr X. tropicalis; Xeru X. ruwenzoriensis; lg X. laevis/X. gilli* laboratory hybrid.

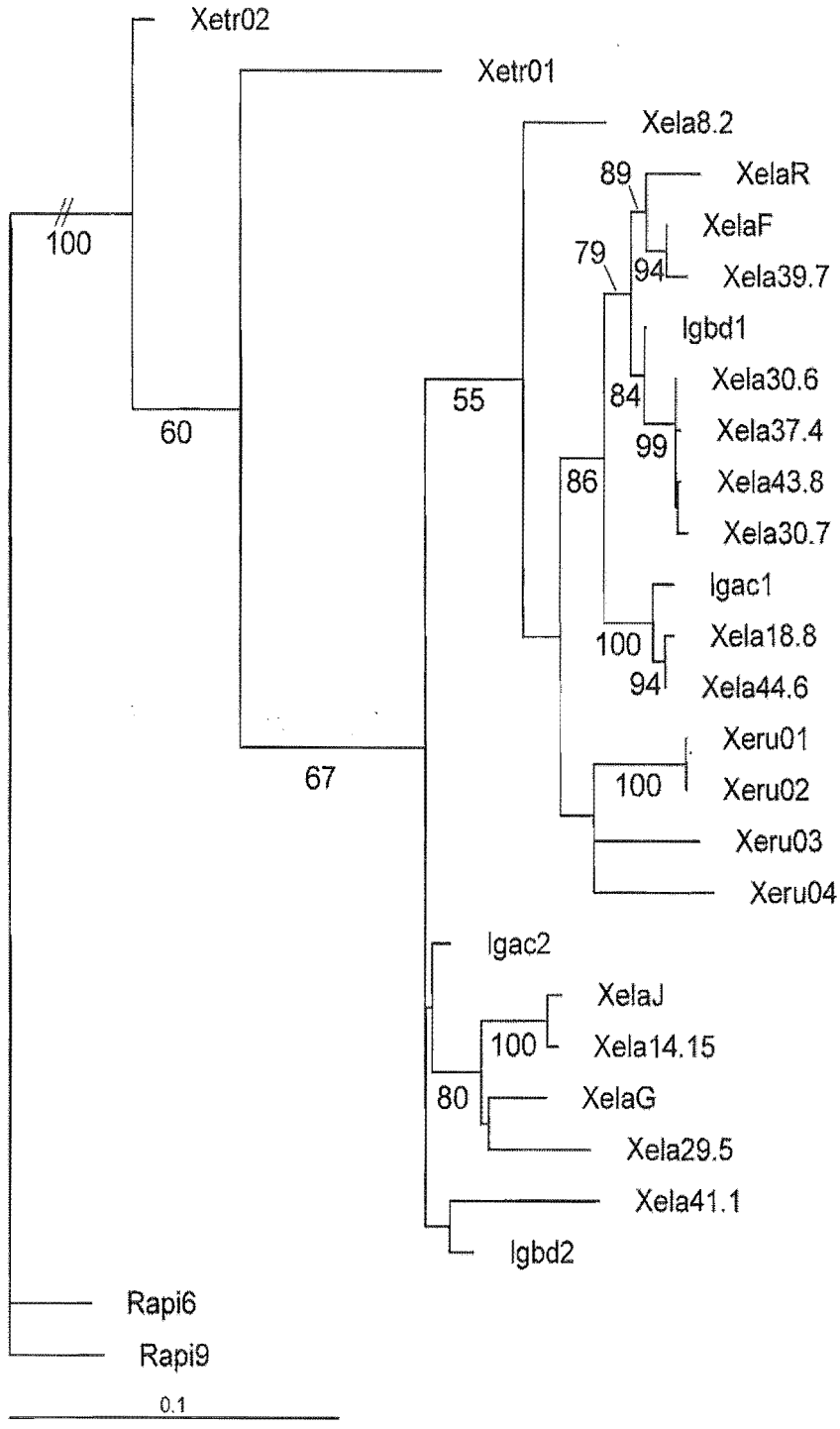


Figure 4.2. Evolutionary relationships of *Xenopus* class Ia sequences with *Rana* outgroup using α2 and α3 domain sequences. Numbers indicate bootstrap support for nodes. All branches scaled as in Figure 4.1.

tropicalis alleles form a sister group to the monophyletic clade consisting of *X. laevis, X. gilli*, and *X. ruwenzoriensis*. This topology establishes the separation of *X. tropicalis* alleles, and monophyly of other sequences of the ingroup. All *X. ruwenzoriensis* alleles form a monophyletic cluster nested within a larger clade of *X. laevis* and *X. gilli*. There is a closely related group that forms a tight cluster similar to that seen in the α1 domain tree, but the cluster is comprised of different sequences and contains only one recombinant allele. This tree also has a mixture of both long and short terminal branches, but compared to the α1 domain tree, branches on the α2/α3 tree are much shorter.

Several a priori hypotheses were compared to the ML and NJ topologies for both the α1 and α2/α3 sequence partitions. *A priori* hypotheses are not exhaustive, but are designed to gauge similarity to various levels of trans-species allelic sharing. Hypotheses included are: 1) reciprocal monophyly of species ((*X. laevis*),(*X. ruwenzoriensis*),(*X.tropicalis*),(*R. pipiens*)), 2) unconstrained *X. laevis* (*X. laevis*, (*X. ruwenzoriensis*),(*X. tropicalis*),(*R. pipiens*)), unconstrained *X. laevis* and *X. ruwenzoriensis* (*X. laevis*, *X. ruwenzoriensis*, (*X. tropicalis*),(*R. pipiens*)), 3) unconstrained *Xenopus* species (*X. laevis*, X. *ruwenzoriensis*, X. *tropicalis* (*R. pipiens*)) and 4) monophyletic *X. laevis* ((*X. laevis*), *X. ruwenzoriensis*, X. *tropicalis*, (*R. pipiens*)). Results of the Shimodaira-Hasegawa tests for both data partitions show that there is no significant difference between the ML and NJ trees (Table 4.3). However, hypotheses 1 and 4 are significantly different from unconstrained optimal trees for both data partitions. For the α1 domain, hypothesis 2 is also significantly different from the optimal tree. For the α2/α3 data partition, tree

topology	lnL	δ lnL	S-H[a]	W S-H[b]
	α1 domain trees			
ML	-1844.420	-	-	-
3	-1844.905	0.5	0.850	0.812
NJ	-1858.748	14.3	0.491	0.281
2	-1920.962	76.5	0.000	0.000
4	-2009.004	164.6	0.000	0.000
1	-2058.469	214.0	0.000	0.000
	α2/α3 domain trees			
ML	-3003.147	-	-	-
3	-3003.382	0.2	0.848	0.861
2	-3003.382	0.2	0.848	0.877
NJ	-3003.710	0.6	0.743	0.727
1	-3148.052	144.9	0.000	0.000
4	-3168.448	165.3	0.000	0.000

[a]Shimodiara-Hasegawa (1999) test
[b]weighted Shimodiara-Hasegawa (2002) test

Table 4.3. Statistical comparison of phylogenetic hypotheses for *Xenopus* class Ia alleles. The likelihood scores of the ML and NJ topologies are compared against numbered topological constraints representing various levels of transspecies allele sharing.

constraints for hypotheses 2 and 3 resulted in the same phylogenetic reconstruction. Only hypothesis 3, where the only constraint is placing *R. pipiens* as an outgroup to *Xenopus*, is not significantly different from the unconstrained ML and NJ topologies for all data partitions.

DISCUSSION

GENETIC EXCHANGE

Intragenic recombination plays a prominent role in creating and maintaining diversity in the MHC class I of *X. laevis*. Genetic exchange events that were detected occurred so that new functional alleles are created in the MHC by generating new PBR arrangements. This kind of genetic exchange creates a clear pattern of reticulate evolution among class I alleles and is characterized by phylogenetic inconsistencies between the α1 domain and sites in the remaining 3' exons of the gene. This pattern would not be expected from *in vitro* chimera formation, where random breakpoints would be expected; therefore I interpret this result to be due to *in vivo* genetic exchange (Judo et al. 1998). Patterns of *X. laevis* recombination are in contrast to the reticulate evolution seen among alleles of humans for 3 class I loci (Jakobsen et al. 1998). In humans, the HLA-B locus displays the strongest signal of genetic exchange, but no single recombination break point is especially prominent this or other class I loci (Hughes et al. 1993).

The prevalence of genetic exchange in *X. laevis* could be the result of large intron size. A similar explanation was proposed for salmonid fishes, and subsequent sequence data confirmed that intron II separating the exons coding the α1 and α2 domains is 2.6 kb in length (Shum et al. 2002). I attempted to PCR amplify the intron spanning exons 2 and 3 of the class Ia gene but was unable to recover any specific amplification product. Although the size of intron II in *X. laevis* at the class Ia locus remains uncertain, introns in class II MHC genes in this species are known to be *ca.* 10- 17 kb in length (Kobari et al. 1995). Also, different classes of repetitive elements have been identified in *X. laevis* (Carroll et al. 1989), so it is probable that the rate of genetic exchange in class I MHC introns is influenced by these genomic features.

Although I have identified one recombination breakpoint in these sequences, the location of the other endpoint remains uncertain. However, class I alleles show strong linkage disequilibrium with alleles of functionally and physically linked LMP and TAP loci (Namikawa et al. 1995; Ohta et al. 2003). This linkage is possible only if the genetic

exchange events occur over relatively small segments of DNA compared to the distance between genetic markers, as can be the case with gene conversions (Andolfatto and Nordborg 1998). The length of the genetic exchange detected here is therefore expected to be much smaller than the distance separating these loci. Although genetic exchanges in *X. laevis* are relatively small on a genomic scale, they are larger and different from those characterized in primates.

PHYLOGENETICS OF XENOPUS MHC

Class I alleles fall into lineages that are long lasting and predate certain speciation events within the genus *Xenopus*. These conclusions are upheld when evolutionary relationships are reconstructed with either α1 domain sequences or α2/α3 domain sequences. One surprising note is that the trees from the different domains are not statistically different. The two data partitions used here differ because of recombination; therefore, several causes for the similarity or correlation between the trees are possible. Similar topologies may arise if recombination events occurred in the distant past, if recombination events occurred between closely related sequences, or if recombination events involved nonreciprocal exchange of genetic material (Posada and Crandall 2002). A lack of statistical difference in trees could also partly reflect the conservative nature of the test or lack of strong phylogenetic signal because the test uses a bootstrapping procedure to measure differences in trees (Shimodaira and Hasegawa 1999). In reconstructing evolutionary analysis on MHC genes, other authors have used the "total evidence" approach (Kluge 1989) to establish the species distribution of allelic lineages (e.g. Shum et al. 2001). Such an approach using ML methods on these data also confirm the above conclusions (data not shown).

Based on phylogenetic results, trans-species sharing of allelic lineages among certain divergent *Xenopus* species takes place. Comparison of this observation with scenarios that constrain allelic separation among various species groups confirms this result. Class Ia trans-species evolution in *Xenopus* extends to species thought to be much more evolutionarily divergent than the species among which trans-species evolution is commonly found in primate class I genes. *Xenopus ruwenzoriensis* and *X. gilli* alleles cluster together in a clade nested within *X. laevis* alleles. *Xenopus tropicalis* samples do not appear to share allelic lineages with other species sampled here. While the class Ia sequences from all *Xenopus* species form a monophyletic clade, *X. tropicalis* diverged

from other *Xenopus* species prior to the formation of the extant class Ia lineages. This is not surprising since the divergence time of *X. tropicalis* species and the *Xenopus* common ancestor is estimated at about 100 million years ago (Graf 1996).

MODE OF CLASS IA MHC EVOLUTION IN X. LAEVIS

MHC class Ia evolution in *X. laevis* is more similar to the mode of MHC evolution found in salmonid fishes than mammals. Teleost fish and *Xenopus* frogs share at least three features of MHC evolution: 1) levels of polymorphism exceeding that found in primates, 2) a distinct pattern of genetic exchange, and 3) class Ia lineages that persist for long periods of time. All three of these features differ from the mode of class Ia evolution found in primates (Shum et al. 2001; Vogel et al. 1999). Also, others have noted that some nonmammalian vertebrates have fewer classical class I loci than is generally found in mammals (Kaufman 1999; Ohta et al. 2002; Trowsdale 1995).

The mode of MHC evolution described in salmonid fishes, and now found in *X. laevis,* may be the result of genomic features of the MHC region. A previous supposition is that the unusual mode of evolution is fish-specific, and potentially caused by the separation of class I and class II loci onto different chromosomes in teleost fish (Shum et al. 2001). If true, one would expect the MHC of *X. laevis* to evolve in a manner more similar to humans because both these species have a single MHC region within which both the class Ia and class II regions are found (Nonaka et al. 1997a). Instead, *X. laevis* and salmonid fishes share a common mode of class Ia MHC evolution, providing evidence that this particular mode of MHC evolution is not primarily due to the separation of the class I and class II regions. Undoubtedly the non-linkage of class I and class II regions in teleost fishes influences the mode of MHC evolution, but I have shown that it is not vital for extended retention of ancestral class Ia lineages in *X. laevis*.

MHC class I processing genes (immunoproteasome components, TAP transporter genes and tapasin) are located close to MHC class Ia genes in *Xenopus* (Namikawa et al. 1995; Ohta et al. 1999). In fishes, these processing genes are also found closely linked to the classical class Ia gene (Takami et al. 1997). In humans and mice however, the class I processing genes are paradoxically found in the class II region rather than in the class I region (Beck et al. 1992). A likely result of this genomic organization is that distinctive allelic associations exist among class Ia and class I processing genes in *Xenopus* frogs and other taxa, but are not known in primates and mice (Joly et al. 1998; Kaufman 1999; Ohta

et al. 2003). Therefore, it is possible that the mode of evolution common to *Xenopus* and fishes is due in part to the number of class Ia loci, linkage and possible co-evolution of this suite of genes. Differences in the class I region may arise due to stronger co-evolutionary tendencies of, or differences in stability of genes in close proximity to the MHC class Ia loci (Amadou 1999; Kaufman 1999).

Other factors may affect the mode of evolution in *Xenopus* and salmonid fishes. For instance, both *Xenopus* and salmonids are tetraploid but only one copy of the MHC class I gene has been detected. After polyploidization, the genomes of resultant species are known to undergo rapid changes that include gene silencing, deletion and genome reorganization (Soltis and Soltis 1995). The dynamic nature of the polyploid genome may also lead to a change in linkage or patterns of recombination in the MHC region, leading to differences in evolution. The observation of different modes of MHC evolution in various taxa is consistent with the duplication, deletion and divergence of loci, but differs in that the timing, frequency, strength and duration of evolutionary events is distinct from that observed in mammals (Edwards et al. 1995). The different mode of evolution in *Xenopus* and fishes may be due to the close linkage of MHC class I and MHC processing genes such as proteasome components and TAP transporter elements. Since this pattern is also seen in sharks, (Ohta et al. 2002; Ohta et al. 2000) it may be widespread among primitive vertebrates, and aspects of this mode of evolution may represent the ancestral mode of evolution.

ACKNOWLEDGEMENTS

Martin Flajnik, and Neil Gemmell provided useful comments on this manuscript and Louis DuPasquier provided ideas and invaluable technical assistance in the laboratory and in the animal breeding and care facility. This research is supported by the Marsden Fund (Royal Society of New Zealand) and a PhD scholarship from the University of Canterbury.

REFERENCES

Abbas AK, Lichtman AH, Pober JS (2000) Cellular and Molecular Immunology. W. B. Saunders, Philadelphia

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control AC 19:716-723

Amadou C (1999) Evolution of the Mhc class I region: the framework hypothesis. Immunogenetics 49:362-367

Andolfatto P, Nordborg M (1998) The effect of gene conversion on intralocus associations. Genetics 148:1397-1399

Beck S, Kelly A, Radley E, Khurshid F, Alderton RP, Trowsdale J (1992) DNA sequence analysis of 66 kb of the Human MHC class II region encoding a cluster of genes for antigen processing. Journal of Molecular Biology 228:433-441

Bingulac-Popovic J, Figueroa F, Sato A, Talbot WS, Johnson SL, Gates M, Postlethwait JH, Klein J (1997) Mapping of MHC class I and class II regions to differnet linkage groups in the zebrafish, *Danio rerio*. Immunogenetics 46:129-134

Bontrop RE, Otting N, de Groot N, Doxiadis GGM (1999) Major histocompatibility complex class II polymorphisms in primates. Immunological reviews 167:339-350

Cadavid LF, Watkins DI (1997) The duplicative nature of the MHC class I genes: an evolutionary perspective. European Journal of Immunogenetics 24:313-322

Carroll D, Knutzon DS, Garrett JE (1989) Transposable elements in *Xenopus* species. In: Berg DE, Howe MM (eds) Mobile DNA. American Society for Microbiology, Washington, D. C.

Edwards SV, Wakeland EK, Potts WK (1995) Contrasting histories of avian and mammalian MHC genes revealed by class II B sequences from songbirds. Proceedings of the National Academy of Sciences, USA 92:12200-12204

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368-376

Figueroa F, Gunther E, Klein J (1988) MHC polymorphism pre-dating speciation. Nature 335:265-267

Flajnik MF, Kaufman J, Riegert P, Du Pasquier L (1984) Identification of class I major histocompatibility complex encoded molecules in the Amphibian *Xenopus*. Immunogenetics 20:134-143

Flajnik MF, Ohta Y, Greenberg AS, Salter-Cid L, Carrizosa A, Du Pasquier L, Kasahara M (1999) Two ancient allelic lineages at the single classical class I locus in the *Xenopus* MHC. Journal of Immunology 163:3826-3833

Graf J-D (1996) Molecular approaches to the phylogeny of *Xenopus*. In: Tinsley RC, Kobel HR (eds) The biology of *Xenopus*. Clarendon Press, Oxford, p 379-389

Hall T (1999) Bioedit: a user-friendly biological sequence alignment editor and analysis program for Window 95/98/NT. Nucl. Acids Symp. Ser. 41:95-98

Hess CM, Edwards SV (2002) The evolution of the major histocompatibility complex in birds. BioScience 52:423-431

Hughes AL, Hughes MK, Watkins DI (1993) Contrasting roles of interallelic recombination at the HLA-A and HLA-B loci. Genetics 133:669-680

Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex loci of vertebrates. Annual Review of Genetics 32:415-435

Jakobsen IB, Wilson SR, Easteal S (1998) Patterns of reticulate evolution for the classical class I and II HLA loci. Immunogenetics 48:312-323

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics 29:217-222

Joly E, Le Rolle AF, Gonzalez AL, Mehling B, Stevens J, Coadwell WJ, Hunig T, Howard JC, Butcher GW (1998) Co-evolution of rat TAP transporters and MHC class I RT1-A molecules. Current Biology 8:169-172

Judo MSB, Wedel AB, Wilson C (1998) Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Research 26:1819-1825

Kaufman J (1999) Co-evolving genes in MHC haplotypes: the "rule" for nonmammalian vertebrates? Immunogenetics 50:228-236

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proceedings of the National Academy of Sciences, USA 78:454-458

Kluge AG (1989) A concern for evidence and a phylogenetic hypothesis of relationships among *Epicurates* (Boidae, Serpentes). Systematic Zoology 38:7-25

Kobari F, Sato K, Shum BP, Tochinai S, Katagiri M, Ishibashi T, Du Pasquier L, Flajnik MF, Kasahara M (1995) Exon-intron organization of *Xenopus* MHC class II B chain genes. Immunogenetics 42:376-385

Lawlor DA, Ward FE, Ennis PD, Jackson AP, Parham P (1988) *HLA-A* and *HLA-B* polymorphism predate the divergence of humans and chimpanzees. Nature 335:268-271

Martin D, Rybicki E (2000) RDP: detection of recombination amongst aligned sequences. Bioinformatics 16:562-563

Martinsohn JT, Sousa AB, Guethlein LA, Howard JC (1999) The gene conversion hypothesis of MHC evolution: a review. Immunogenetics 50:168-200

Namikawa C, Salter-Cid L, Flajnik MF, Kato Y, Nonaka M, Sasaki M (1995) Isolation of Xenopus *LMP-7* homologues: striking allelic diversity and linkage to MHC. Journal of Immunology 155:1964-1971

Nonaka M, Namikawa C, Kato Y, Sasaki M, Salter-Cid L, Flajnik MF (1997) Major histocompatibility complex gene mapping in the amphibian *Xenopus* implies a primordial organization. Proceedings of the National Academy of Sciences, USA 94:5789-5791

Ohta Y, McKinney EC, Criscitiello MF, Flajnik MF (2002) Proteosome, Transporter associated with antigen processing, and class I genes in the Nurse Shark *Ginglymostoma cirratum:* evidence for a stable class I region and MHC haplotype lineages. Journal of Immunology 168:771-781

Ohta Y, Okamura K, McKinney EC, Bartl S, Hashimoto K, Flajnik MF (2000) Primitive synteny of vertebrate histocompatibility complex class I and class II genes. Proceedings of the National Academy of Sciences, USA 97:4712-4717

Ohta Y, Powis SJ, Coadwell WJ, Haliniewski DE, Liu Y, Li H, Flajnik MF (1999) Identification and mapping of *Xenopus* TAP2 genes. Immunogenetics 49:171-182

Ohta Y, Powis SJ, Lohr RL, Nonaka M, Du Pasquier L, Flajnik MF (2003) Two highly divergent ancient allelic lineages of the transporter associated with antigen processing (TAP) gene in *Xenopus*: further evidence for co-evolution among MHC class I region genes. European Journal of Immunology 33:3017-3027

Parham P (1994) The rise and fall of great class I genes. Seminars in Immunology 6:373-382

Parham P, Ohta T (1996) Population biology of antigen presentation by MHC class I molecules. Science 272:67-74

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818

Posada D, Crandall KA (2001a) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proceedings of the National Academy of Sciences, USA 98:13757-13762

Posada D, Crandall KA (2001b) Intraspecific gene genealogies: trees grafting into networks. Trends in Ecology and Evolution 16:37-45

Posada D, Crandall KA (2001c) Selecting models of nucleotide substitution: an application to Human Immunodeficiency Virus (HIV-1). Molecular Biology and Evolution 18:897-906

Posada D, Crandall KA (2002) The effect of recombination on the accuracy of phylogeny estimation. Journal of Molecular Evolution 54:396-402

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4:406-425

Sambrook E, Fritsch F, Maniatis T (1989) Molecular Cloning. Cold Spring Harbor Press, Cold Spring, NY

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. Systematic Biology 51:492-508

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Molecular Biology and Evolution 16:1114-1116

Shimodaira H, Hasegawa M (2001) CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246-1247

Shum BP, Avila D, Du Pasquier L, Kasahara M, Flajnik MF (1993) Isolation of a classical MHC class I cDNA from an amphibian. Evidence for only one class I locus in the *Xenopus* MHC. Journal of Immunology 151:5376-5386

Shum BP, Guethlein LA, Flodin LR, Adkinson MA, Hedrick RP, Nehring RB, Stet RJM, Secombes C, Parham P (2001) Modes of Salmon MHC class I and II evolution differ from the primate paradigm. Journal of Immunology 166:3297-3308

Shum BP, Mason PM, Magor KE, Flodin LR, Stet RJM, Parham P (2002) Structures of two major histocompatibility complex class I genes of the rainbow trout (*Oncorhynchus mykiss*). Immunogenetics 54:193-199

Soltis DE, Soltis PS (1995) The dynamic nature of polyploid genomes. Proceedings of the National Academy of Sciences, USA 92:8089-8091

Swofford DL (1998) PAUP* Phylogenetic Analysis Using Parsimony (*and other methods). Version 4.0. Sinauer, Sunderland, MA

Takami K, Zaleska-Rutczynska Z, Figueroa F, Klein J (1997) Linkages of *LMP, TAP,* and *RING3* with *Mhc* class I rather than class II genes in the Zebrafish. Journal of Immunology 159:6052-6060

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 22:4673-4680

Trowsdale J (1995) Both man and bird and beast: comparative organization of MHC genes. Immunogenetics 41:1-17

Vogel TU, Evans DT, Urvater JA, O'Connor DH, Hughes AL, Watkins DI (1999) Major histocompatibility complex class I genes in primates: co-evolution with pathogens. Immunological reviews 167:327-337

Xia X, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. Molecular Phylogenetics and Evolution 26:1-7

CHAPTER V

PRECAUTION USING CONFORMATIONAL GENOTYPING METHODS ON MHC CLASS I GENES

ABSTRACT

MHC genes are among the most diverse loci in the vertebrate genome, and the diversity is thought to be adaptively maintained by balancing selection. As a result no single allele at a locus is driven to fixation, but instead many alleles are found in high frequency in a population. Because of these characteristics, MHC genes are popularly used for conservation genetics and as a potential recognition allele locus in behavioural ecology. When using MHC genes for these purposes, often many individuals in a population must be screened and a rapid method of genotyping is used. Because of the unique intron-exon structure of the class I gene, problems may arise with the use of conformational genotyping techniques to characterize variation in a class I locus. Results indicate that single-strand conformational polymorphism (SSCP) genotyping of MHC class I exon 2 sequences underestimates the number of MHC alleles and misspecifies alleles as identical when they are different. The consequences of these shortcomings are discussed in light of conservation and behavioural experiment design.

INTRODUCTION

Major histocompatibility complex (MHC) genes encode proteins involved in the cellular response of the vertebrate immune system. There are two types of MHC proteins, the class I and class II, and while they have similar structure and function, differences also exist. Class I proteins are expressed on almost all cells and present peptides derived from the internal cellular environment to CD^+ 8 T-cells (Abbas et al. 2000). Class II proteins are expressed mainly on specialized antigen-presenting cells of the immune system, and present peptides derived from extracellular proteins to CD^+ 4 helper T-cells. In both cases, the MHC-bound peptides are recognized as either foreign antigens or as a native component of the body. Cells expressing foreign peptides are targeted for destruction and once infected cells are identified, a cascade of signals activates other components of the immune system.

Genes encoding MHC proteins are highly polymorphic and have high levels of allelic diversity at the population level (Parham and Ohta 1996). The polymorphism of MHC genes is widespread among taxa and is thought to be due to the operation of some form of natural selection (Hughes and Yeager 1998). As a result, there is no wild-type MHC allele in a population, but instead many allelic variants are found at high frequencies. Also, natural selection acts to maintain some allelic lineages of MHC genes and persistence times of MHC lineages extend beyond the formation of closely related species (Takahata and Nei 1990; Takahata et al. 1992). Because of the extensive variation and diversity of adaptive MHC genes, few individuals in a population share an MHC allele unless they are closely related. These characteristics have made the study of MHC genes attractive not only in immunology, but to evolutionary and behavioural biology as well (Edwards and Hedrick 1998).

When collecting MHC gene data in evolutionary or behavioural biology, large numbers of individuals are often assayed. Experiments or observations in these fields involve hundreds of samples, so that genotyping or collecting allelic information is very time consuming. Therefore, quick genotyping methods that rely on conformational changes in DNA folding and gel electrophoresis are used rather than sequencing alleles. Some of these methods are single strand conformation polymorphisms (SSCP) and denaturing gradient gel electrophoresis (DGGE), which are popular because they are fast, relatively inexpensive and often do not require specialized equipment (Orita et al. 1989).

Another advantage is that these methods typically do not require extensive characterization prior to genotyping, enabling the use of conformational techniques in nonmodel organisms (Edwards et al. 2000).

Because conformational genotyping can be done on large numbers of samples using equipment already commonly found in the lab, this method is frequently used in population genetics in conjunction with an MHC gene. For example, the allelic diversity in a population or among closely related species has been measured with conformational techniques (e.g. Kim et al. 1999; Pfau et al. 1999; e.g. Van der Walt et al. 2001). MHC genes are specifically used because of the adaptive value of this protein in relation to the fitness of an individual and population. As such, the level of variation at MHC loci can be compared to that of neutrally evolving (e.g. microsatellite) loci using SSCP or DGGE. These types of studies can also be extended to comparisons across different environments or geographical ranges within a species (Landry and Bernatchez 2001). Frequently, these studies use conformational variants to screen for alleles in the population, and then a representative sample of each different band is sequenced. This minimizes the numbers of samples that need to be sequenced to obtain data on all alleles in a data set (e.g. Garrigan and Hedrick 2001). Though frequently used in evolutionary biology, SSCP or DGGE genotyping of MHC genes is of particular interest in animals with small population sizes.

The use of conformational techniques and MHC genes in conservation biology is becoming commonplace (Girman 1996). Frequently endangered populations are genotyped at neutrally evolving loci, but the MHC provides an adaptive gene locus the directly impacts fitness and represents more functional differences between populations (Crandall et al. 2000). Such differences include the variation in susceptibility to pathogens and disease and have strong implications for long-term population survival (Langefors et al. 2001). SSCP or DGGE has been used on MHC genes to genotype animals from a variety of taxa, from fishes to mammals (Garrigan and Hedrick 2001; Richman et al. 2001). Because of the adaptive value of MHC genes, these methods are used to recommend the designation of evolutionary significant units (ESUs) or management units (MUs) in endangered species (e.g. Hedrick et al. 2001). In addition to evolutionary and conservation biology, SSCP and DGGE are used on the MHC in behavioural biology as well.

In behavioural biology, the theory of inclusive fitness indicates that altruism or self-defeating cooperative behaviour could have adaptive value and arise among social

animals if these behaviours are limited to close genetic relatives (Hamilton 1964). The conundrum of inclusive fitness is how animals accurately identify individuals, and several mechanisms have been proposed (Blaustein 1983; Waldman 1987). Among these proposed mechanisms are recognition alleles. These are encoded by a single locus that conveys a signal of kinship for other conspecifics to perceive. Recognition alleles in a population may become wide spread if they accurately allow individuals to identify kin (Waldman 1988). The only way this mechanism can be effective is if the recognition locus had many alleles at the population level; then unrelated individuals would share an allele only infrequently, leading to the accurate identification of kin. In vertebrates the only known genetic locus identified with levels of allelic diversity high enough for use as recognition alleles is that of the MHC (Parham and Ohta 1996).

MHC genes have been used in a variety of behavioural contexts and conformational techniques are often used to assay variation at putative recognition loci (Brown and Eklund 1994). Mate choice and sexual selection experiments help elucidate how animals choose their mates, presumably with the use of MHC to identify partners. To genotype animals used in these investigations, conformational techniques are sometimes used (Landry et al. 2001; Reusch et al. 2001). In the context of mating behaviour, extra-pair matings also can be detected using MHC loci and conformational techniques (Sommer and Tichy 1999). Research into kin discrimination by animals in terms of inclusive fitness also employ these methods (Olsen et al. 2002). Much of this behavioural research in nonmodel organisms has been done on fishes, but the dynamics of mammalian mate choice and kin discrimination also have been investigated (Clarke and Faulkes 1999).

In preparation for behavioural experiments on tadpoles of the African clawed frog *Xenopus laevis*, DNA sequences were obtained from a captive adult breeding population. Offspring of the breeding population were to be used in kin recognition experiments investigating the role of the MHC locus as a component of the recognition mechanism (Landry et al. 2001; Olsen et al. 1998). Because many tadpoles were to be used for experiments, genotyping of the MHC from stimulus groups and experimental subjects was to be done through SSCP. The effectiveness of SSCP depends partly on the length of the fragment used to genotype the individual, with an optimal size of about 200 bp (Sunnucks et al. 2000). I used exon 2 of the class I MHC to characterize the MHC genotype of tadpoles because that segment is the right size for the technique and because the variation in that exon is expected to accurately differentiate alleles at that locus (Hedrick 1994).

However, unlike most investigations using conformational techniques, I sequenced areas outside of the gene segment used for SSCP. Results from sequencing and SSCP analysis contradict each other however, due to the unique patterns of variation of the MHC class Ia gene in *X. laevis*.

MATERIALS AND METHODS

RNA ISOLATION AND SEQUENCING

To genotype adult frogs of *X. laevis*, I took blood samples and washed them in PBS isotonic for amphibians to remove blood plasma and proteins. TRIzol reagent (Life Technologies) was added to washed cells, which were stored at –80° C. I extracted RNA from samples using the TRIzol reagents, following manufacturer's recommendations (Life Technologies). First strand synthesis was performed on total RNA as described in chapter IV. Immediately following first strand synthesis, PCR was performed using primers designed to amplify the α1, α2 and α3 domains of the MHC class I gene. I cloned PCR products into plasmid vectors and transformed into *Escherichia coli* cells, which were grown following methods described in Chapter IV. Sequences were obtained from clones using an ABI 3700 automated sequencer, and editing of trace files was done using Bioedit (Hall 1999).

DNA ISOLATION AND GENOTYPING

DNA was obtained from a small tail clipping from tadpoles that were to be used for behavioural experiments. DNA extraction and purification from samples was performed using a cell-lysis procedure followed by protein precipitation using salt solutions (Fetzner 1999). DNA was precipitated using ethanol, lyophilized, and resuspended in TE buffer. PCR of DNA was performed using 25µl reactions consisting of final concentrations of 0.2mM dntp, 1x buffer, 2.5 mM $MgCl_2$, 0.28µM primers and 1 unit of *Taq* polymerase (Roche). Thirty PCR cycles consisted of denaturing for 0:15 min. at 94 °C, annealing for 0:15 min. at 56 °C, and extension for 0:30 min. at 72 °C. Each PCR run was preceded by an initial denaturing stage at 94° C for 2:00 min. and followed by a 2:00 min. extension stage. SSCP analysis (Orita et al. 1989) was done by adding a standard formamide loading dye (Sambrook et al. 1989) to PCR products and denaturing for 5:00 min. at 94 °C. Following denaturing, samples were immediately incubated in an ice bath for 10:00 min. Four microlitres of sample was loaded on to a 10% nondenaturing polyacrylamide

gel (40% 19 : 1 acrylamide : bis-acrylamide and 10% glycerol) and run at 20 V for 19 hours at room temperature. Bands on the gel were stained in a bath containing 1X SYBR gold staining solution and visualized on a 2400 nm wavelength light box.

RESULTS

Eighteen new sequences from *X. laevis* were obtained through cloned rtPCR techniques. Sequences contain conserved characteristics such as N-glycosation sites, salt bridges and conserved residues consistent with classical class Ia genes from other taxa (Bartl 1998; Flajnik et al. 1991). Amino acid polymorphism in these sequences are also primarily found in sites involved in the putative PBR of the protein(Flajnik et al. 1999). Analysis of nucleotide diversity along different exons of the sequence indicates that exon 2, comprising the α1 domain, has the highest level of diversity, in both of the putative lineages (Table 5.1). Exon 3 amino acid diversity reveals a different pattern, where lineage B has expected levels of diversity based on patterns of polymorphism seen in other species, but lineage A has very low levels of polymorphism (Table 5.1). Including *X. laevis* sequences downloaded from Genbank, lineage A comprises 12 sequences, but when comparing the exon 3 segment of the sequence, only 8 alleles are differentiated. Lineage B is comprised of 10 sequences and has 10 different exon 3 variants. The average number of nucleotide differences between sequences within a lineage also illustrates the differing patterns of polymorphism found in exon 3 of these lineages (Table 5.1). As previously discovered by Flajnik et al. (1999), exon 4 of both lineages has very low levels of diversity despite the high polymorphism seen in other exons of the sequence.

Patterns of SSCP variation indicates certain individuals scored as sharing MHC alleles. However, the sequencing done indicates that sharing of polymorphism between these individuals applies only to exon 3; these individuals have different MHC class Ia alleles when one considers the entire length of the gene that was sequenced. Based on

		exon 2			exon 3			exon 4		
	N	π	K	Hap	π	K	Hap	π	K	Hap
Lineage A	12	0.131	33.26	12	0.023	6	8	0.013	3.20	7
Lineage B	10	0.152	38.49	10	0.103	28	10	0.028	7.13	8

Table 5.1. Results of MHC class I variation by exon. Sequences of the MHC class Ia gene of *X. laevis* were obtained. Sequences were putatively classified into predefined lineages (Flajnik et al. 1999) based on exon 2 sequence, and the number of sequences (N), nucleotide diversity (π), average number of pairwise nucleotide differences (K) and number of haplotypes (Hap) are given to show differences in levels of variation among lineages and exons.

evidence from MHC sequence of *X. laevis*, SSCP banding patterns of one lineage will correctly identify only about 66% of different alleles, and other alleles will be scored as identical. Behavioural tests for which the SSCP genotyping is to be used is designed to examine MHC-based kin discrimination and experimental design requires that stimulus groups and test animals have known genotypes at the MHC. Using these genotypes, tests are done on groups that share none, one or both MHC alleles, but use of SSCP genotyping on MHC class I exon 3 will result in identifying groups that are scored as sharing one or both alleles, but actually have differing alleles, thus nullifying results of the behavioural tests.

DISCUSSION

The use of conformational genotyping methods can be efficient, cost-effective and a time-saver. The problems associated with using the method here do not stem from limitations of the method, but are instead a result of genetic architecture of the MHC class I gene. The highly polymorphic peptide binding region of the MHC class I protein is comprised of amino acids encoded from both exon 2 and exon 3 of the MHC class I α chain (Bjorkman et al. 1987b). In humans, exon 2 differences will discriminate all but 2.4% of pairwise comparisons of alleles at *HLA-B*. Exon 3 will accurately predict individual alleles of the entire gene in all but 1.8% of those same comparisons (Parham et al. 1995). Based on that information, and the knowledge that class I proteins are conserved in structure and function among taxa (Kaufman et al. 1994), I designed SSCP genotyping in *X. laevis* on exon 3 sequences to get maximum resolution of alleles.

Patterns of variation in each exon are generally seen in *X. laevis*, but for several animals, there is an unexpected low level of allelic diversity in exon 3 (Table 5.1). When SSCP genotyping is used, sequencing of different bands is usually done to check that different bands have different sequences (e.g. Landry et al. 2001). On the other hand, co-migrating bands are assumed to be identical and are not checked to see if they are different. In this case such sequencing detected that some identical bands are indeed identical in the fragment used for SSCP, but are part of different alleles when the entire PBR coding sequence is accounted for. In *X. laevis* sequences, identical SSCP exon 3 bands are found among several different alleles, and false positive identification of individuals that share MHC alleles is common. The phenomenon of spuriously identifying identical alleles using these methods can also easily occur in humans and mice, where

identical class Ia exons exist among different alleles (Johnson and Wu 1998). These results highlight one potential shortcoming of the SSCP method, and are different in nature from SSCP banding patterns that are unable to differentiate distinct alleles that differ at sites within the SSCP fragment (Orita et al. 1989; Reusch et al. 2001).

While SSCP banding patterns will likely accurately reflect allelic diversity for the fragment used for analysis, sometimes important variation exists in the gene of study that is outside the fragment used for SSCP analysis. These differences will affect results of experiments that rely on allelic assignments made by genotyping methods. Similar problems could be encountered using other types of genetic markers such as DGGE and RFLP, or partial sequencing of a gene. Here I have sequenced roughly 800 nucleotides of the MHC class I gene, comprising exons 2-4, but four more short exons exist in the gene and additional variation may exist in these regions. When genotyping individuals for behavioural or ecological applications, care should be taken as to which method is used, and information pertaining to the levels of variation in the gene and the how it is distributed among functional domains of the protein should also be considered when choosing a method. The intron-exon structure of the functional domains may also affect the patterns of variation seen when using incomplete fragments of genes as surrogates to identify alleles.

Often MHC genes are used in behavioural or conservation studies because of the variation found in the MHC genes. For behavioural studies the potential of the MHC as a recognition allele locus is attractive and warrants examination. In conservation biology, the MHC is attractive because variation is thought to be adaptively maintained (Apanius et al. 1997) and information at this locus provides additional information next to variation at genes that evolve neutrally (Crandall et al. 2000). Few examples of adaptive genes are available, so MHC gene sequencing for conservation purposes is becoming common. Care should be taken when interpreting results however, because limitations as to the length of conformational fragment can allow some variation to go unnoticed.

The best way to avoid the possibility of poor inference or design of studies due to undetected variation in the MHC is to assay variation at multiple loci. One study in which several polymorphic loci are assayed is that of Liu et al. (2002). In that paper, they investigate disease resistance in young chickens to *Salmonella enteritidis*. This study benefits from the well known genomic organization and structure of the chicken MHC region. They assay all known polymorphic domains of various loci, both in the class I and

class II genes using conformational techniques (illustrated in Figure 5.1B). Such a complete study is rare and is difficult to do in organisms that are not model organisms, where little is known about the numbers and location of polymorphic class I and class II genes. In another example, the polymorphism at the MHC is assayed at single class I and class II loci in whitefish (Binz et al. 2001). Although multiple class I loci are known, but not surveyed for variation, polymorphism at genes linked to all known variable MHC regions have been detailed (see Figure 5.1C). In both cases, conclusions of the study regarding the effects of MHC polymorphisms to *Salmonella* resistance and level of genetic variation in whitefish are generally robust and complete. These types of studies serve as excellent examples of research with accurate information regarding the affect of MHC polymorphism and a phenotype of interest.

In an interesting behaviour study in stickleback fishes, a correlation between mate choice and numbers of shared alleles was found (Reusch et al. 2001). Here, multiple loci were assayed, but they were all class II loci (Figure 5.1D). An interesting correlation was found in that females choose mates to maximize the number of different alleles at the tested loci, but do not engage in disassortative mating *per se*. However, we do not know the level of variation at the polymorphic class I domains or if they correlate with some sort of mating preference. Although this study represents an important step forward in behaviour studies by genotyping multiple loci through conformational techniques, an incomplete picture is painted because we are missing information on several polymorphic domains. We might draw different conclusions regarding the mate choice behaviour of sticklebacks if class I polymorphism was surveyed and did show a correlation with disassortive mating.

In an example that is more typical of the kind of genotyping that is performed most often in behaviour and disease resistance studies, Arkush et al. (2002) test resistance to three pathogens (including a virus and a bacterium) in Chinook salmon. They use a conformational technique to genotype the fish at a single class II MHC locus (see example in Figure 5.1E). They neither found an effect of specific genotypes to bacterial infection nor a difference in survivability between heterozygotes and homozygotes to the bacterial pathogen. They did however find a higher level of survival among heterozygotes when challenged with a viral infection. Because for most infections the entire immune system works together and both class I and class II molecules can bind and display both bacterial and viral peptides, any classical MHC locus could influence the course of infection. In

this study the authors do find that a gene unlinked to class II is having an effect on viral infection and resistance. This could very well be the class I locus and not genotyping at this locus is a significant shortcoming in this and many other studies (e.g. Langefors et al., 2001; Van Der Walt et al., 2001; Richman et al., 2001; Sommer and Tichy, 1999). Very different conclusions might be drawn if more complete information were available, such as a specific class I allele that offers advantage in viral infection. Also, a correlation with heterozyosity or total number of alleles among several class II loci may have been missed because they were not assayed.

Similar sorts of MHC screening (involving only a single locus) commonly occurs in behaviour studies. Using SSCP to genotype individuals at the single class II locus of Atlantic salmon, Landry et al. (2001) test how mate choice correlates with MHC variation, but have no data on the class I polymorphic domains. They found that mate choice occurs to maximize the differences at the PBR of the class II locus, but not to minimize the number of shared alleles. However, if a class I locus was typed a different result could be envisioned and would open up the possibility that mating occurred to minimize the number of shared alleles (as in Reuch et al. 2001) or that matings correlate with amino acid differences.

In a similar study (Wedekind, et al. 2004), female mate choice is tested and correlated with variation at a single MHC locus. Results indicate no evidence for an effect of class II variation and fusion of gametes. These authors do indicate that class I and class II genes may work together in many regards, but do not test if class I can correlate with fusion of gametes. Clearly a different outcome might be drawn if class I variation were surveyed.

A final example represents yet another step in the decreasing completeness in the number of polymorphic domains assayed for a particular phenotype. In this research only one of the two polymorphic domains of a single class I locus is assayed in Chinook salmon (Garrigan and Hedrick, 2001). Such a screening of polymorphism is likely an inadequate representation of the variation found at this gene locus and the MHC in general (Figure 5.1F). For instance, gene conversion can create several alleles that share variable sites in one domain but are markedly different in other domains. Such variation as well as all variation at other polymorphic loci is entirely missed in such a sampling design. A more complete sampling of potentially polymorphic domains can significantly alter conclusions drawn at to the amount of adaptive variation found in a population, potential

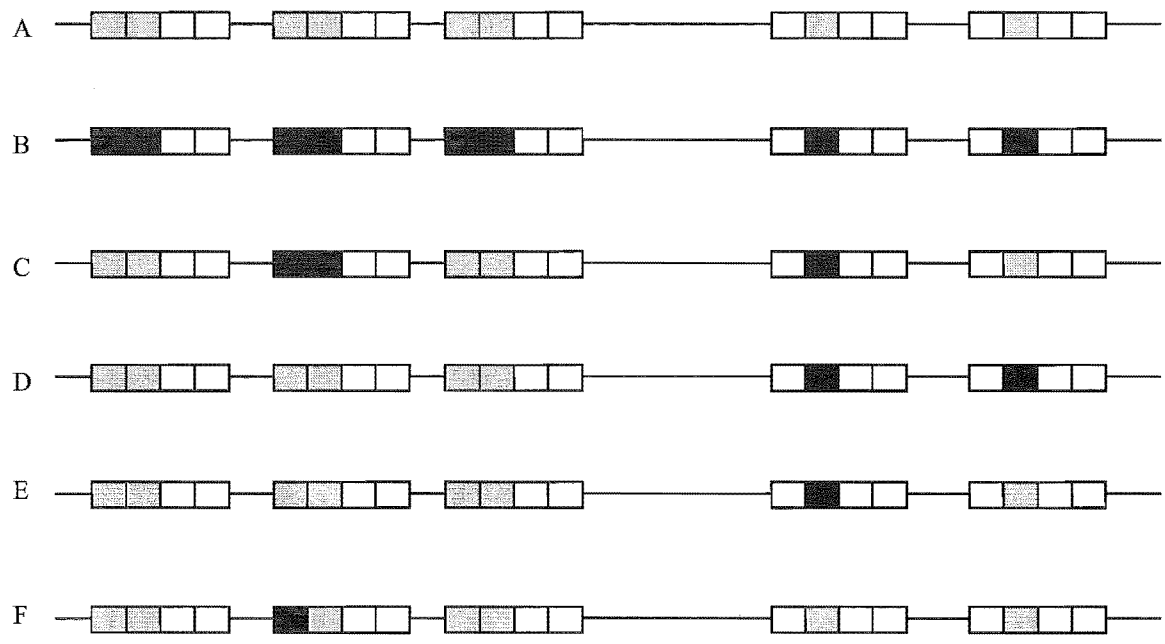


Figure 5.1. The figure above illustrates various scenarios of sampling of polymorphic domains of MHC genes. In A, the MHC of a hypothetical organism is shown in which three class I loci, each with two polymorphic domains shaded in grey, and two class II loci, each with a single polymorphic domain, are located in a region spanning tens or hundreds of kilobases. In various scenarios, polymorphic domains that are assayed for variability are shaded in black. In an ideal scenario visualized in B, all known polymorphic domains are surveyed for variation. This can only be done in model organisms in which all MHC gene are known and which have relatively low numbers of class I and class II loci. Otherwise, subsets of polymorphic domains are assayed. In C, a class I and class II gene is sampled, and this polymorphism serves as a surrogate for variation in the class I and class II regions. This typically is an adequate strategy because multiple class I genes are usually found in haplotypes inherited as a block, and multiple class II loci are as well, but class I and class II loci are often found out of linkage disequilibrium. In D, multiple genes in either the class I or class II region are assayed, but levels of variation in the other region is unsampled. E represents a common scenario in which a single class II gene is assayed for variation, and polymorphism at all other loci is unknown. Finally, F represents a situation in which only one of the two polymorphic domains of a class I locus is sampled for variation.

for susceptibility to disease, and long term survival probabilities. Before recommending courses of management action, or concluding a (lack of) certain types of correlation with MHC variation and behaviours or disease resistance, an adequate screening the immune system variation should be made.

The MHC class I genes are particularly prone to some of the limitations mentioned above because of the intron-exon structure of the gene. In MHC class II proteins, the PBR is also the location of most of the variation and differences here are responsible for operational differences of alleles and ligand binding (Brown et al. 1993; Hughes et al. 1994). In class II genes, the part of the β chain gene that encodes the PBR residues are found on a single exon, so that almost all of the variation of the gene is found at that exon. Also, the gene encoding the α chain of the protein does encode part of the PBR, but the α chain typically has little or no variation at the functional level. Thus SSCP analysis done on MHC class II genes is less likely to suffer from the problems identified with class I genes of *X. laevis*.

Each MHC class I and class II gene is usually one of several members of a multigene family, several loci of which may play an important role in the immune system (Beck and Trowsdale 2000). Often when using a gene fragment for behaviour or conservation, interest is focused solely on that single locus. However, many loci of class I and class II genes exist in most species (Trowsdale 1995) and variation at other loci is often ignored. One exception has been noted in which several class II loci were genotyped in the exon 2 encoding part of the PBR (Reusch et al. 2001). Even this study falls short however, because any highly variable MHC locus is a candidate as a recognition allele locus, but no polymorphic class I loci were genotyped. That study was done in fish, and since class I and class II genes in fishes are unlinked (Bingulac-Popovic et al. 1997; Sato et al. 2000), we cannot expect variation at one class to correlate with alleles at the other class of MHC genes. Similarly, in conservation genetics, a single locus may be genotyped and inference may be made regarding low variation at that locus, implying that the immune system is somehow deficient (O'Brien and Evermann 1988; Yuhki and O'Brien 1990). However, low variation at a single locus does not mean that the entire adaptive immune system is deficient (Sanjayan et al. 1996), and variation at more than one locus should be tested before there is cause for alarm (O'Brien 1994). The immune system is a complex part of the body, and care should be taken when using just one part of a gene in or an incomplete fragment of a gene, as the pattern of variation at that locus may not be representative of other loci or other exons.

REFERENCES

Abbas AK, Lichtman AH, Pober JS (2000) Cellular and Molecular Immunology. W. B. Saunders, Philadelphia

Apanius V, Penn DJ, Slev P, Ruff LR, Potts WK (1997) The nature of selection on the Major Histocompatibility Complex. CRC Critical Reviews in Immunology 17:179-224

Arkush KD, Giese A, Mendonca HL, McBride AM, Marty GD, Hedrick PW (2002) Resistance to three pathogens in the endangered winter-run chinook salmon (*Oncorhynchus tshawytscha*): effects of inbreeding and major histocompatibility complex genotypes. Canadian Journal of Fisheries and Aquatic Science 59:966-975

Bartl S (1998) What sharks can tell us about the evolution of MHC genes. Immunological Reviews 166:317-331

Beck S, Trowsdale J (2000) The human Major Histocompatibility Complex: lessons from the DNA sequence. Annual Review of Genomics and Human Genetics 1:117-137

Bingulac-Popovic J, Figueroa F, Sato A, Talbot WS, Johnson SL, Gates M, Postlethwait JH, Klein J (1997) Mapping of MHC class I and class II regions to differnet linkage groups in the zebrafish, *Danio rerio*. Immunogenetics 46:129-134

Binz T, Largiader C, Muller R, Wedekind C (2001) Sequence diversity of MHC genes in lake whitefish. Journal of Fish Biology 58:359-373

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC (1987) Structure of the human class I histomcompatibility antigen, HLA-A2. Nature 327:506-512

Blaustein AR (1983) Kin recognition mechanisms: phenotypic matching or recognition alleles. American Naturalist 121:749-754

Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 364:33-39

Brown JL, Eklund A (1994) Kin recognition and the major histompatibility complex: an intergrative review. American Naturalist 143:435-461

Clarke FM, Faulkes CG (1999) Kin discrimination and female mate choice in the naked mole rat *Heterocephalus glaber*. Proceedings of the Royal Society of London 266:1995-2002

Crandall KA, Bininda-Edmonds ORP, Mace GM, Wayne RK (2000) Considering evolutionary processes in conservation biology: an alternative to "Evolutionary significant units". Trends in Ecology and Evolution 15:290-295

Edwards SV, Hedrick PW (1998) Evolution and ecology of MHC molecules: from genomics to sexual selection. Trends in Ecology and Evolution 13:305-311

Edwards SV, Nusser J, Gasper J (2000) Characterization and evolution of Major Histocompatibility Complex (MHC) genes in non-model organisms, with examples from birds. In: Baker AJ (ed) Molecular Methods in Ecology. Blackwell Science, Cambridge, p 168-205

Fetzner JW (1999) Extracting high quality DNA from shed reptile skins: a simplified method. Biotechniques 26:1052-1054

Flajnik MF, Canel C, Kramer J, Kasahara M (1991) Evolution of the major histocompatibility complex: molecular cloning of major histocompatibility complex class I from the amphibian *Xenopus*. Proceedings of the National Academy of Sciences, USA 88:537-541

Flajnik MF, Ohta Y, Greenberg AS, Salter-Cid L, Carrizosa A, Du Pasquier L, Kasahara M (1999) Two ancient allelic lineages at the single classical class I locus in the *Xenopus* MHC. Journal of Immunology 163:3826-3833

Garrigan D, Hedrick PW (2001) Class I MHC polymorphism and evolution in endangered California Chinook and other Pacific salmon. Immunogenetics 53:483-489

Girman D (1996) The use of PCR-based single strand conformation polymorphism analysis (SSCP-PCR) in conservation genetics. In: Smith TB, Wayne RK (eds) Molecular Genetic approaches in Conservation. Oxford University Press, Oxford, p 167-182

Hall T (1999) Bioedit: a user-friendly biological sequence alignment editor and analysis program for Window 95/98/NT. Nucl. Acids Symp. Ser. 41:95-98

Hamilton WD (1964) The genetical evolution of social behaviour. I. Journal of Theoretical Biology 7:1-16

Hedrick PW (1994) Evolutionary genetics of the major histocompatibility complex. American Naturalist 143:945-964

Hedrick PW, Parker KM, Lee RN (2001) Using microsatellite and MHC variation to identify species, ESUs and MUs in the endangered Sonoran Topminnow. Molecular Ecology 10:1399-1412

Hughes AL, Hughes MK, Howell CY, Nei M (1994) Natural selection at the class II major histocompatibility complex loci in mammals. Philosophical Transcripts of the Royal Society of London 345:359-367

Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex loci of vertebrates. Annual Review of Genetics 32:415-435

Johnson G, Wu TT (1998) Possible Assortment of a1 and a2 Region Gene Segments in Human MHC Class I Molecules. Genetics 149:1063-1067

Kaufman J, Salomonsen J, Flajnik MF (1994) Evolutionary conservation of MHC class I and class II molecules--different yet the same. Seminars in Immunology 6:411-424

Kim T, Parker KM, Hedrick PW (1999) Major histocompatibility complex differentiation in Sacramento River Chinook Salmon. Genetics 151:1115-1122

Landry C, Bernatchez L (2001) Comparative analysis of population structure across environments and geographical scales at major histocompatibility complex and microsatellite loci in Atlantic Salmon (*Salmo salar*). Molecular Ecology 10:2525-2539

Landry C, Garant D, Duchesne P, Bernatchez L (2001) 'Good genes as heterozygostiy': the major histocompatibility complex and mate choice in Atlantic salmon (*Salmo salar*). Proceedings of the Royal Society of London 268:1279-1285

Langefors A, Lohm J, Grahn M, Anderson O, Von Schantz T (2001) Association between major histocompatibility complex class IIB alleles and resistance to *Aeromonas salmonicida* in Atlantic salmon. Proceedings of the Royal Society of London 268:479-485

Liu W, Miller MM, Lamont SJ (2002) Association of MHC class I and class II polymorphisms with vaccine or challenge response to *Salmonella enteritidis* in young chicks. Immunogenetics 54:582-590

O'Brien SJ (1994) The cheetah's conservation controversy. Conservation Biology 8:1153-1155

O'Brien SJ, Evermann JF (1988) Interactive influence of infectious disease and genetic diversity in natural populations. Trends in Ecology and Evolution 3:254-259

Olsen KH, Grahn M, Lohm J (2002) Influence of MHC on sibling discrimination in Arctic charr, *Salvelinus alpinus* (L.). Journal of Chemical Ecology 28:783-795

Olsen KH, Grahn M, Lohm J, Langefors A (1998) MHC and kin discrimination in juvenile Arctic charr, *Salvelinus alpinus* (L.). Animal Behaviour 56:319-327

Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proceedings of the National Academy of Sciences, USA 86:2766-2770

Orti G, Hare MP, Avise JC (1997) Detection and isolation of nuclear haplotypes by PCR-SSCP. Molecular Ecology 6:575-580

Parham P, Adams EJ, Arnett KL (1995) The origins of HLA-A,B,C polymorphism. Immunological reviews 143:141-180

Parham P, Ohta T (1996) Population biology of antigen presentation by MHC class I molecules. Science 272:67-74

Pfau RS, Van Den Bussche RA, McBee K, Lochmiller RL (1999) Allelic diversity at the *Mhc-DQA* locus in cotton rats (*Sigmodon hispidus*) and a comparison of *DQA* sequences within the family Muridae (Mammalia: Rodentia). Immunogenetics 49:886-893

Reusch TBH, Haberli MA, Aeschlimann PB, Milinski M (2001) Female sticklebacks count alleles in a strategy of sexual selection explaining MHC polymorphism. Nature 414:300-302

Richman AD, Herrera LG, Nash D (2001) MHC class II beta sequence diversity in the deer mouse (*Peromyscus maniculatus*): implications for models of balancing selection. Molecular Ecology 10:2765-2773

Sambrook E, Fritsch F, Maniatis T (1989) Molecular Cloning. Cold Spring Harbor Press, Cold Spring, NY

Sanjayan MA, Crooks K, Zegers G, Foran D (1996) Genetic variation and the immune response in natural populations of pocket gophers. Conservation Biology 10:1519-1527

Sato A, Figueroa F, Murray BW, Malaga-Trillo E, Zaleska-Rutczynska Z, Sultmann H, Toyosawa S, Wedekind C, Klein J (2000) Nonlinkage of major histocompatibility complex class I and class II loci in bony fishes. Immunogenetics 51:108-116

Sommer SC, Tichy H (1999) Major histocompatibility complex (MHC) class II polymorphism and paternity in the monogamous *Hypogeomys antimena*, the endangered, largest endemic Malagasy rodent. Molecular Ecology 8:1259-1272

Sunnucks P (2000) Efficient genetic markers for population biology. Trends in Ecology and Evolution 15:199-203

Sunnucks P, Wilson ACC, Beheregaray LB, Zenger K, French J, Taylor AC (2000) SSCP is not so difficult: the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. Molecular Ecology 9:1699-1710

Takahata N, Nei M (1990) Allelic genealogy under overdominant and frequency dependent selection and polymorphism of major histocompatibility complex loci. Genetics 124:967-978

Takahata N, Satta Y, Klein J (1992) Polymorphism and balancing selection at Major Histocompatibility Complex loci. Genetics 130:925-938

Trowsdale J (1995) Both man and bird and beast: comparative organization of MHC genes. Immunogenetics 41:1-17

Van der Walt JM, Nel LH, Hoelzel AR (2001) Characterization of major histocompatibility complex DRB diversity in the endemic South African antelope *Damaliscus pygarus*: a comparison in two subspecies with different demographic histories. Molecular Ecology 10:1679-1688

Waldman B (1987) Mechanisms of kin recognition. Journal of Theoretical Biology 128:159-185

Waldman B (1988) The ecology of kin recognition. Annual review of ecology and systematics 19:543-571

Wedekind C, Walker M, Portmann J, Cenni B, Muller R, Binz T (2004) MHC-linked susceptibility to a bacterial infection, but no MHC-linked cryptic female choice in whitefish. Journal of Evolutionary Biology 17:11-18

Yuhki N, O'Brien SJ (1990) DNA variation of the mammalian major histocompatibility complex reflects genomic diversity and population history. Proceedings of the National Academy of Sciences, USA 87:836-840

SUMMARY

INTRODUCTION AND AIMS

The immune system is a complex array of tissues, cells, proteins and other molecules that interact at many different levels. Similarly, proteins involved in the immune response have a complex evolutionary history (Parham and Ohta 1996). Shaped by natural selection and genetic drift which are manifest through molecular mechanisms such as mutation and genetic exchange, these proteins have complicated patterns of genetic variation (Yeager and Hughes 1999). Study of the function, structure and variation of genes of the immunes system have been ongoing, but comparative studies of the immune system in basal taxa are relatively new (Flajnik 1998). While progress has been made in understanding the evolutionary dynamics shaping immuno-genes, many questions remain unanswered. Because comparative immunogenetics is still relatively new discipline, there are many gaps in knowledge, but some problems remain unanswered because of inadequacies of methodologies. The MHC class Ia, genes and the evolution of proteasome components are examples of areas for which questions remain unanswered.

The purpose of this thesis is to investigate molecular evolutionary parameters in the *X. laevis* MHC class Ia gene and to explore the dynamics of evolutionary sequence divergence in *psmb5* and *lmp7*. The approach involves the use of newer phylogenetic model-based methods to study, in order to overcome problems associated with shortcomings of older analysis techniques. Data collection involved the cloning and sequencing a population sample of MHC class Ia alleles from *X. laevis* and bioinformatic assembly of *psmb5* and *lmp7* genes of various taxa from the Genbank database. Because all methods used in this thesis rely on phylogenetic information, Chapter I deals with efficiency and accuracy of phylogenetic methods and aims to demonstrate the usefulness of model-based phylogenetic inference. In a similar fashion, Chapter III demonstrates the advantages of evolutionary model-based inference at the population level by estimating substitution rates in MHC alleles under conditions for which pair wise estimators failed. In Chapter II, the dynamics of differentiation after a gene duplication event are studied in *psmb5* and *lmp7* to elucidate patterns of substitutions, and investigate reasons for protein divergence. The mode of MHC class Ia evolution in *X. laevis* is established and the underlying causes of different modes of MHC evolution are clarified in Chapter IV.

Finally, Chapter V outlines difficulties associated with MHC class Ia conformational genotyping techniques.

RESULTS

In Chapter I, results show that model-based methods offer flexible and efficient solutions to phylogenetic problems that arise because of high levels of DNA variation. A traditional way of dealing with sequences that are substitutionally saturated is to remove data from consideration. This solution decreases the size of the sample by making the sequences shorter, and as a result, increases potential bias in parameter estimates. Also, use of overly-simplistic models can lead to systematic error for samples that highly divergent and do not conform to the molecular clock (Bruno and Halpern 1999; Rzhetsky and Sitnikova 1996). Instead, using model-based phylogenetic methods with models that are selected using objective, statistically rigorous criteria can improve results by accurately estimating substitutions in the data (Swofford et al. 1996). Accurate estimation of tree topologies allows better approximation when using molecular evolutionary analysis that relies on phylogenetic information.

Gene duplications and the questions of causes of elevated substitution rates are examined using phylogenetic methods in Chapter II. Previous investigation indicate that the rate of evolution is elevated following the duplication of *psmb5* and *lmp7*, but no evidence for natural selection was detected (Takezaki et al. 2002). Results of Chapter II contradict these finding and conclude that differing substitution rates along different branches of a phylogenetic topology show that natural selection increased the substitution rates immediately following duplication. Discrepancies between these and previous findings can be explained by the use of methods that detect periodic directional selection at specific codons among different lineages in these analyses (Yang and Nielsen 2002). These findings imply that the subfunctionalization may be the driving force behind the increased rate of evolution in *lmp7*.

The purpose of Chapter III was to estimate substitution rate parameters in highly polymorphic alleles and characterize polymorphisms among class Ia alleles in *X. laevis*. In these data traditional pair wise methods failed to accurately estimate substitution rates due to high allelic polymorphism (Flajnik et al. 1999; Shum et al. 2001). Here, results estimate substitution rates in these data and show that nonsynonymous substitution rates are elevated in codons of the peptide binding region in MHC class Ia of salmonid fishes

and *Xenopus* frogs. These patterns are consistent with expectations for the operation balancing selection at this locus. Data from this chapter also elucidates patterns of polymorphism and provides clues to the molecular basis for the strong linkage disequilibrium among functionally linked genes. Also, these data show how the different lineages in *X. laevis* may be functionally different and bind different sets of peptides.

I address the mode of MHC class Ia evolution in *X. laevis* in Chapter IV. Hypotheses from previous work indicate that the mode of MHC evolution characterized in salmonid fishes may be due to the non-linkage of the class I and class II regions in the fish genome (Shum et al. 2001). Based on this work, the mode of MHC class Ia evolution in *X. laevis* is expected to be more similar to mammals than fishes because the class I and class II regions are linked. The findings of Chapter V indicate that patterns of genetic exchange in *X. laevis* are dominated by breakpoints in intron II, and typically involve tracts that are longer than 200 nucleotides. This pattern of recombinations lead to shuffling of PBR domains and is similar in nature to allelic recombination seen in salmonid fishes, but different to patterns seen in mammals (Jakobsen et al. 1998; Shum et al. 2001).

Phylogenetic analysis of MHC class Ia alleles from various frog species reveals that samples from different species cluster together by allelic lineage, rather than by species. This pattern can be explained by ancient origin of the locus and long persistence of allelic lineages with low levels of interlocus recombination (Hughes et al. 1993; Vogel et al. 1999). In the past, class Ia MHC allelic lineages were considered more prone to turnover from the birth-death process of gene duplication and as a result, are more dynamic in nature (Cadavid and Watkins 1997). This research stands in contrast to the traditional paradigm, indicating stability of class Ia lineages, and, along with data from salmonid fishes, indicates that the proto-class Ia locus may have been stable rather than dynamic. Finally, these data indicate that the mode of class Ia MHC evolution is more similar to the fish model. These findings contradict the theory that this mode of evolution is primarily due to nonlinkage of class I and class II MHC regions in the genome. Instead this mode of evolution may be more influenced by linkage of class Ia MHC and functionally related processing genes, or copy number of classical class I loci.

In Chapter V, implications for the genotyping of animals at the class I locus using conformational techniques are appraised. Use of conformational techniques such as single strand conformation polymorphism (SSCP) is common in behavioural and conservation

studies. The methods are typically fast, reliable and inexpensive. Use of SSCP is often followed by confirming the sequences of differentially migrating bands, but co-migrating bands are assumed to be identical. I confirmed both different and co-migrating bands by sequencing the exon used for SSCP and adjacent exons. Results indicate that while accurate, SSCP genotyping done on exon 2 class I alleles, may miss important variation and that some co-migrating bands are different in exon 1. This is because of the intron-exon structure of class I loci, where the highly polymorphic PBR is comprised of two exons (Bjorkman et al. 1987b). In class II loci, SSCP is expected to perform much better because variation is typically concentrated in a single exon (Brown et al. 1993).

CONCLUSION

Using approaches detailed earlier, I have examined molecular evolutionary dynamics of the MHC class Ia gene of *X. laevis*, and a phylogenetic sample of the *psmb5/lmp7* gene pair. Results from analysis in Chapters II and IV support the conclusion that model-based statistical genetic procedures offer significant advantages when working with highly polymorphic data sets both at the phylogenetic and population sampling levels. Appropriate use of models improves efficiency and accuracy when estimating phylogenetic trees and when estimating branch lengths, genetic distance measures or substitution rates. Using new methods to analyze the *psmb5/lmp7* gene pair, I document for the first time that nonsynonymous substitutions are elevated in *lmp7* following gene duplication, and conclude that sequence divergence occurred through directional selection. Although other authors have concluded that sequence substitutions have not contributed to the functional divergence of LMP7, evidence for natural selection and asymmetric divergence imply that the rate of evolution is adaptively supported due to the function of the protein. Finally I conclude that the mode of MHC evolution is different from mammals and more similar to other primitive species. This mode of evolution is not primarily due to the nonlinkage of the class I and class II MHC regions as previously hypothesized.

REFERENCES

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC (1987) Structure of the human class I histomcompatibility antigen, HLA-A2. Nature 327:506-512

Brown JH, Jardetzky TS, Gorga JC, Stern LJ, Urban RG, Strominger JL, Wiley DC (1993) Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. Nature 364:33-39

Bruno WJ, Halpern AL (1999) Topological bias and inconsistency of Maximum likelihood using wrong models. Molecular Biology and Evolution 16:564-566

Cadavid LF, Watkins DI (1997) The duplicative nature of the MHC class I genes: an evolutionary perspective. European Journal of Immunogenetics 24:313-322

Flajnik MF (1998) Churchill and the immune system of ectothermic vertebrates. Immunological Reviews 166:5-14

Flajnik MF, Ohta Y, Greenberg AS, Salter-Cid L, Carrizosa A, Du Pasquier L, Kasahara M (1999) Two ancient allelic lineages at the single classical class I locus in the *Xenopus* MHC. Journal of Immunology 163:3826-3833

Hughes AL, Hughes MK, Watkins DI (1993) Contrasting roles of interallelic recombination at the HLA-A and HLA-B loci. Genetics 133:669-680

Hughes AL, Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-170

Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex loci of vertebrates. Annual Review of Genetics 32:415-435

Jakobsen IB, Wilson SR, Easteal S (1998) Patterns of reticulate evolution for the classical class I and II HLA loci. Immunogenetics 48:312-323

Parham P, Ohta T (1996) Population biology of antigen presentation by MHC class I molecules. Science 272:67-74

Rzhetsky A, Sitnikova T (1996) When is it safe to use an oversimplified substitution model in tree making? Molecular Biology and Evolution 13:1255-1265

Shum BP, Guethlein LA, Flodin LR, Adkinson MA, Hedrick RP, Nehring RB, Stet RJM, Secombes C, Parham P (2001) Modes of Salmon MHC class I and II evolution differ from the primate paradigm. Journal of Immunology 166:3297-3308

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK (eds) Molecular Systematics. Sinauer, Sunderland, MA

Takezaki N, Zaleska-Rutczynska Z, Figueroa F (2002) Sequencing of amphioxus *PSMB5/8* gene and phylogenetic position of agnathan sequences. Gene 282:179-187

Vogel TU, Evans DT, Urvater JA, O'Connor DH, Hughes AL, Watkins DI (1999) Major histocompatibility complex class I genes in primates: co-evolution with pathogens. Immunological reviews 167:327-337

Yang Z (2001) Adaptive Molecular Evolution. In: Balding DJ, Cannings C, Bishop M (eds) Handbook of Statistical Genetics. John Wiley and Sons, New York, p 327-350

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends in Ecology and Evolution 15:496-502

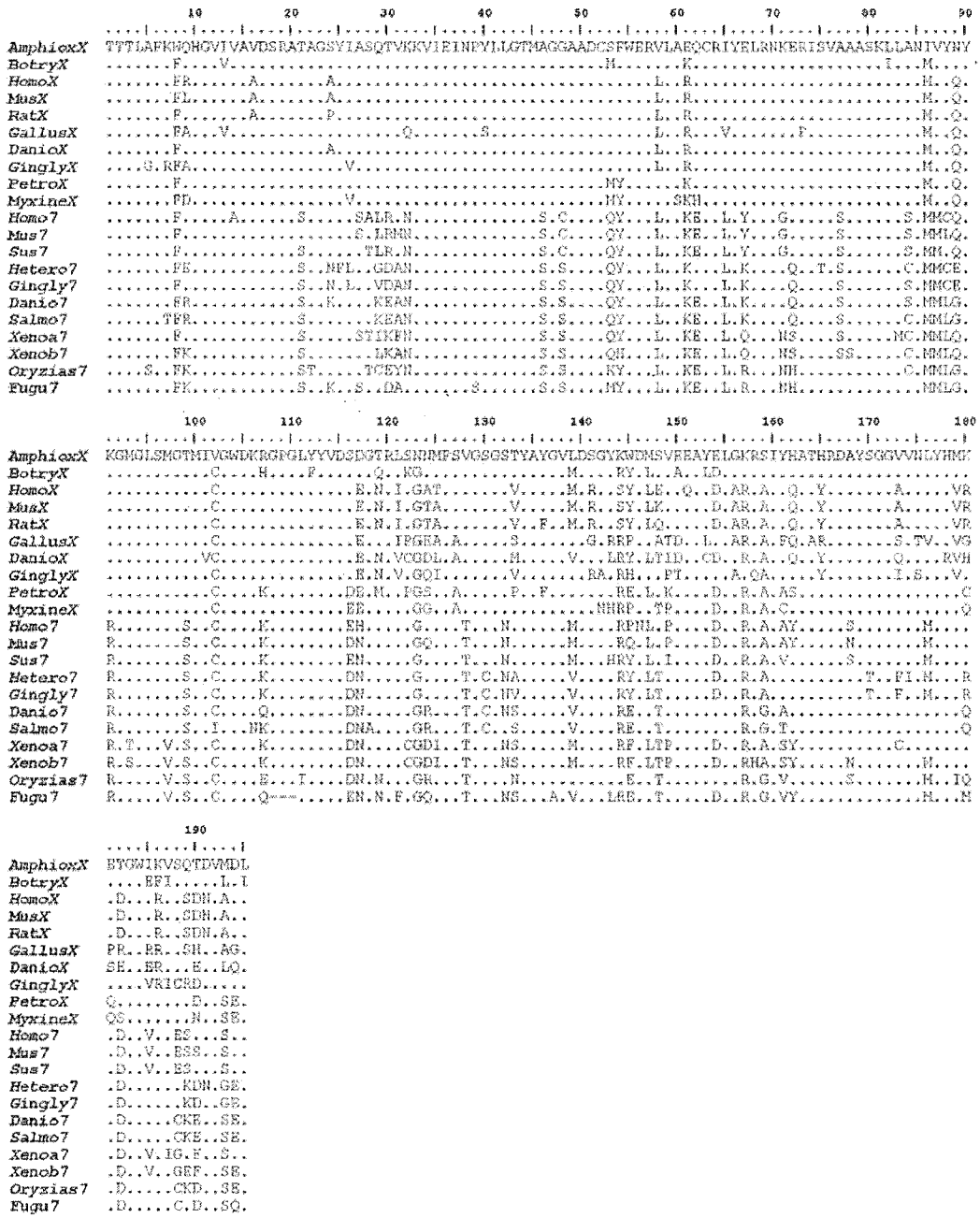
Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution 19:908-917

Yeager M, Hughes AL (1999) Evolution of the mammalian MHC: natural selection, recombination, and convergent evolution. Immunological reviews 167:45-58

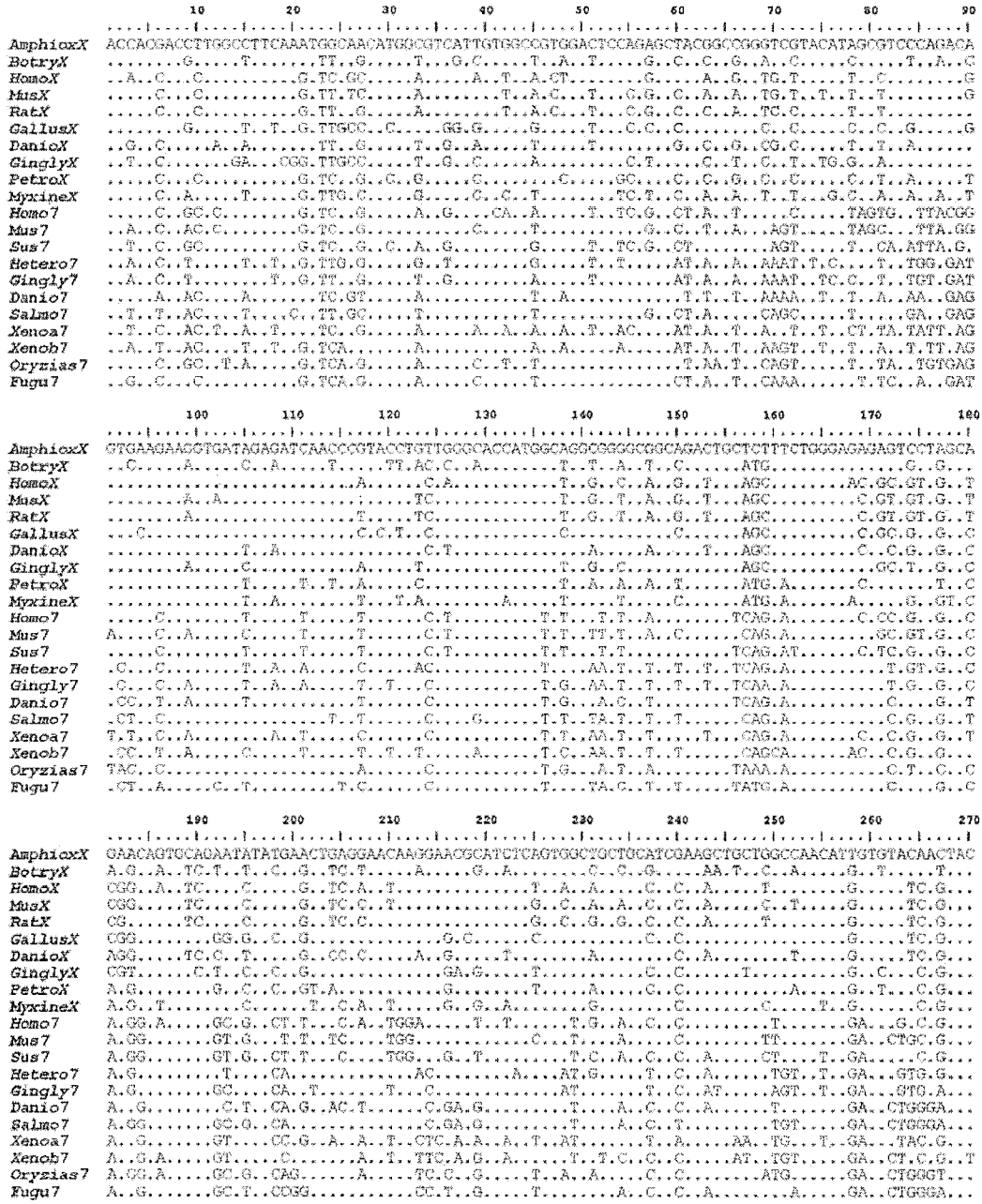
APPENDIX 1

DNA AND PROTEIN SEQUENCE DATA

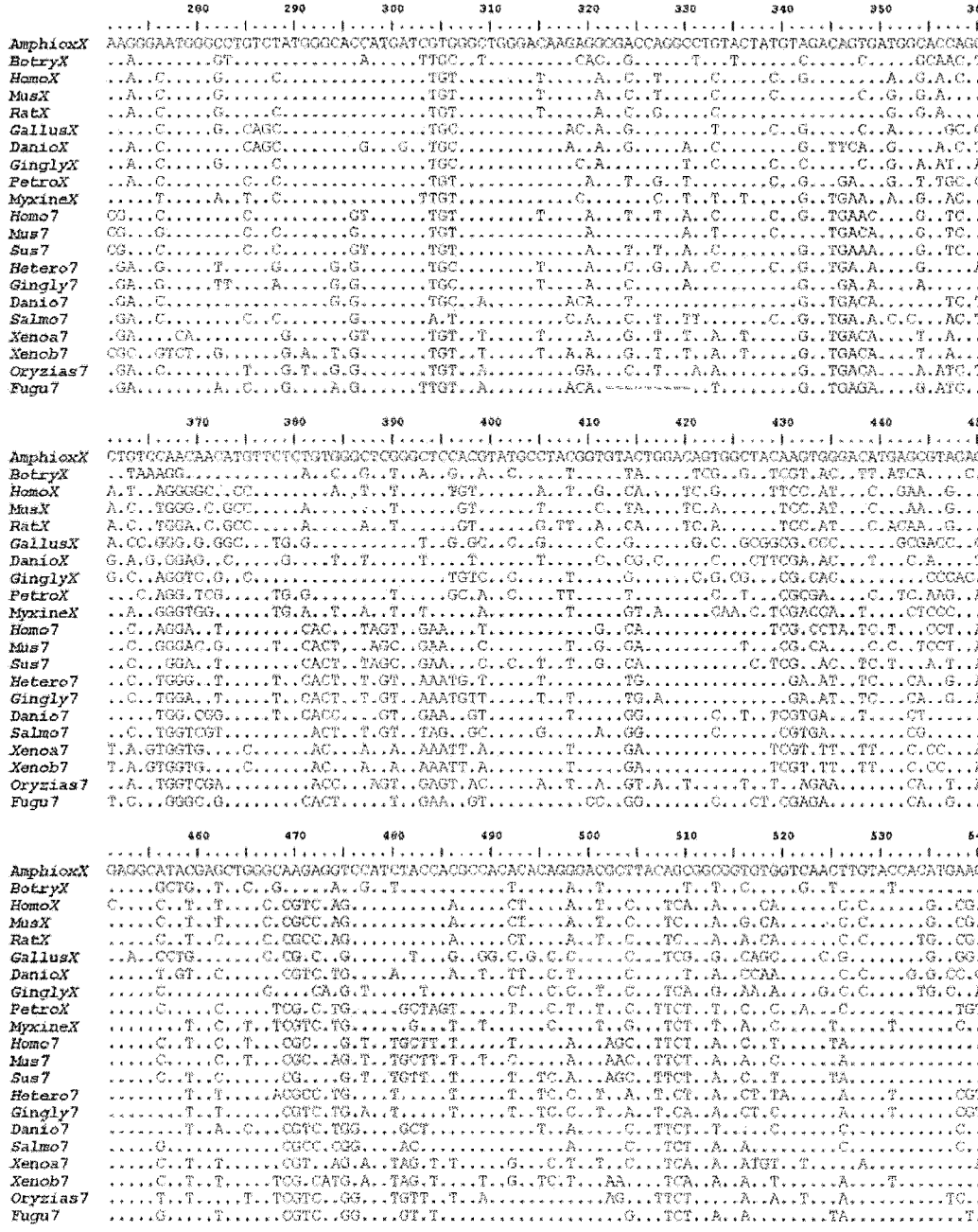
LMP7 and PSMB5 Amino Acid Sequence Data



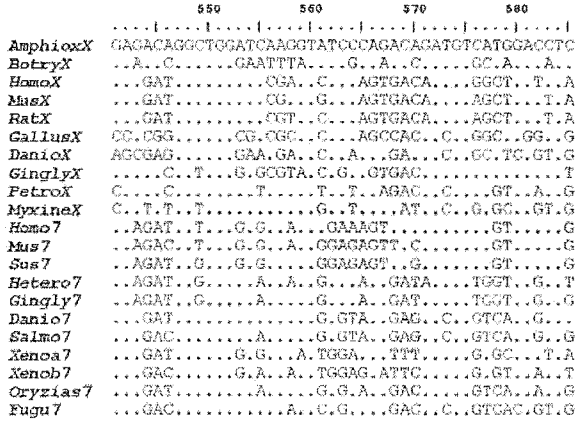
LMP7 and PSMB5 Nucleotide Sequence Data



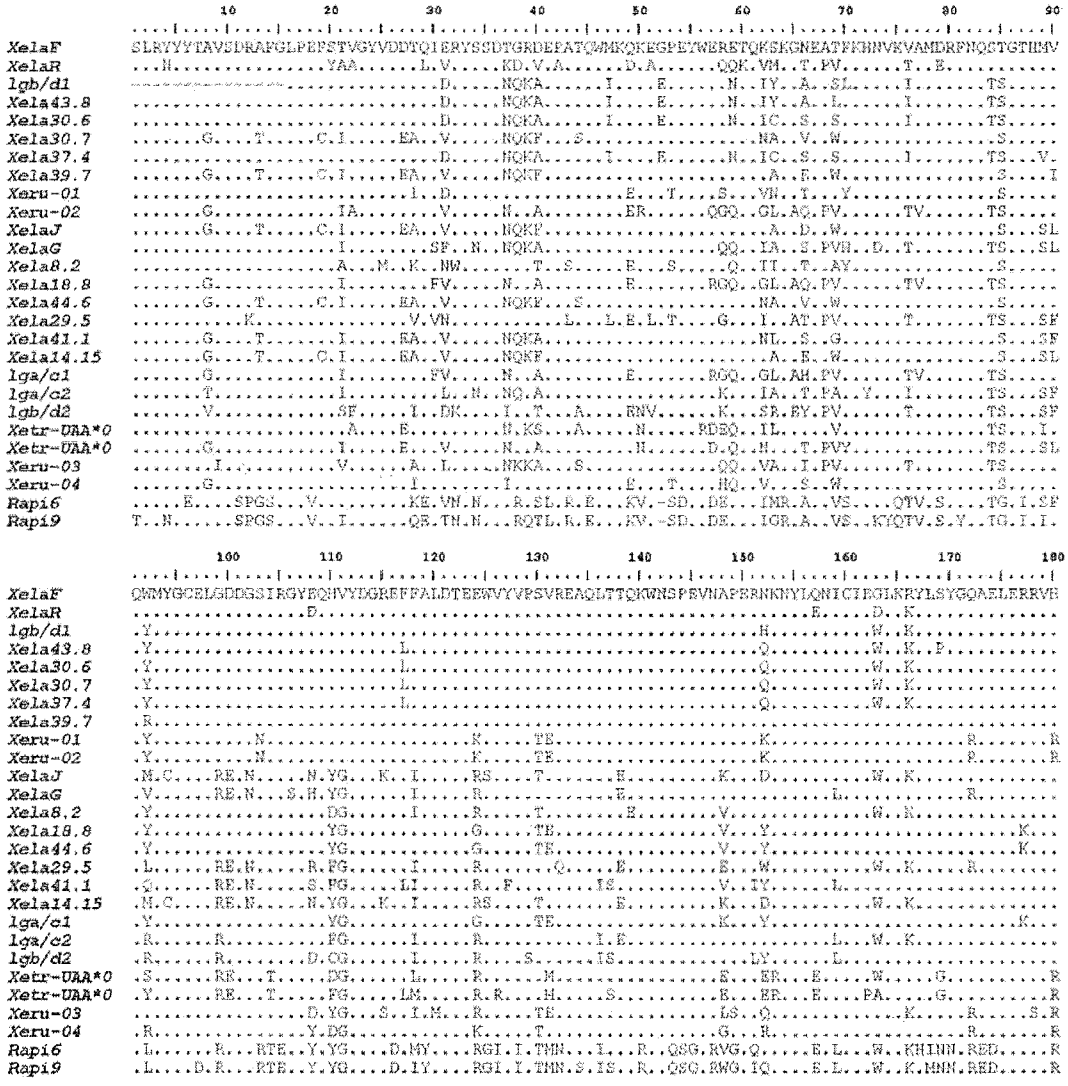
LMP7 and PSMB5 Nucleotide Sequence Data



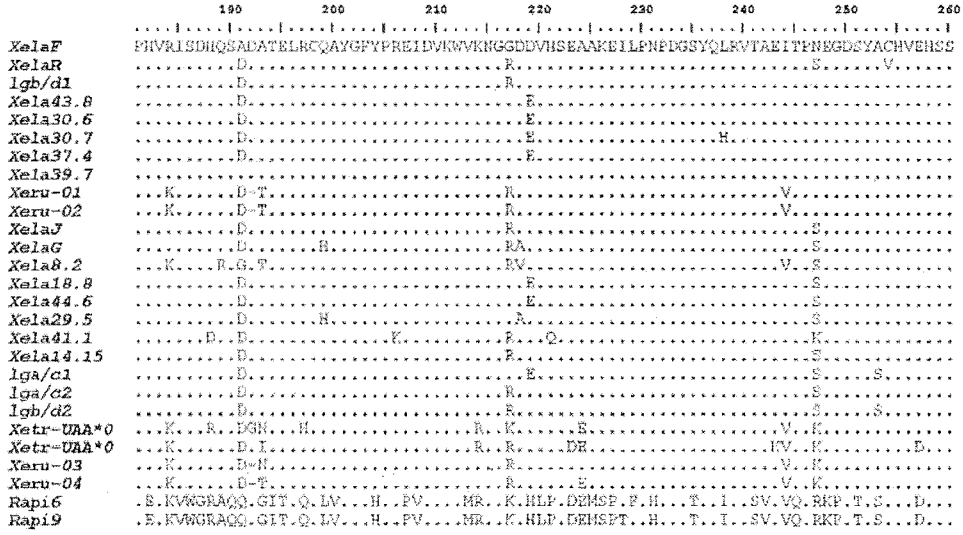
LMP7 and PSMB5 Nucleotide Sequence Data



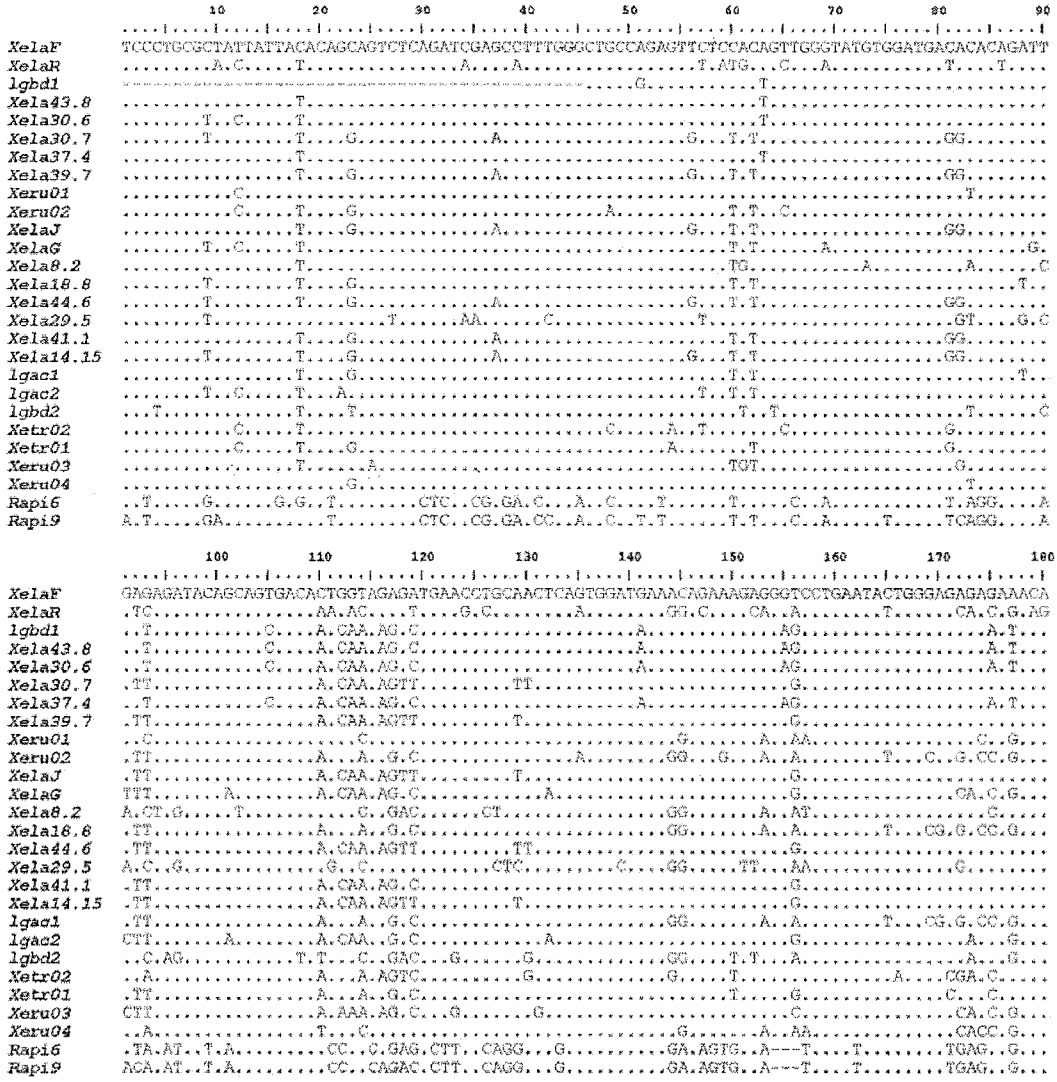
Frog MHC Class Ia Amino Acid Sequence Data



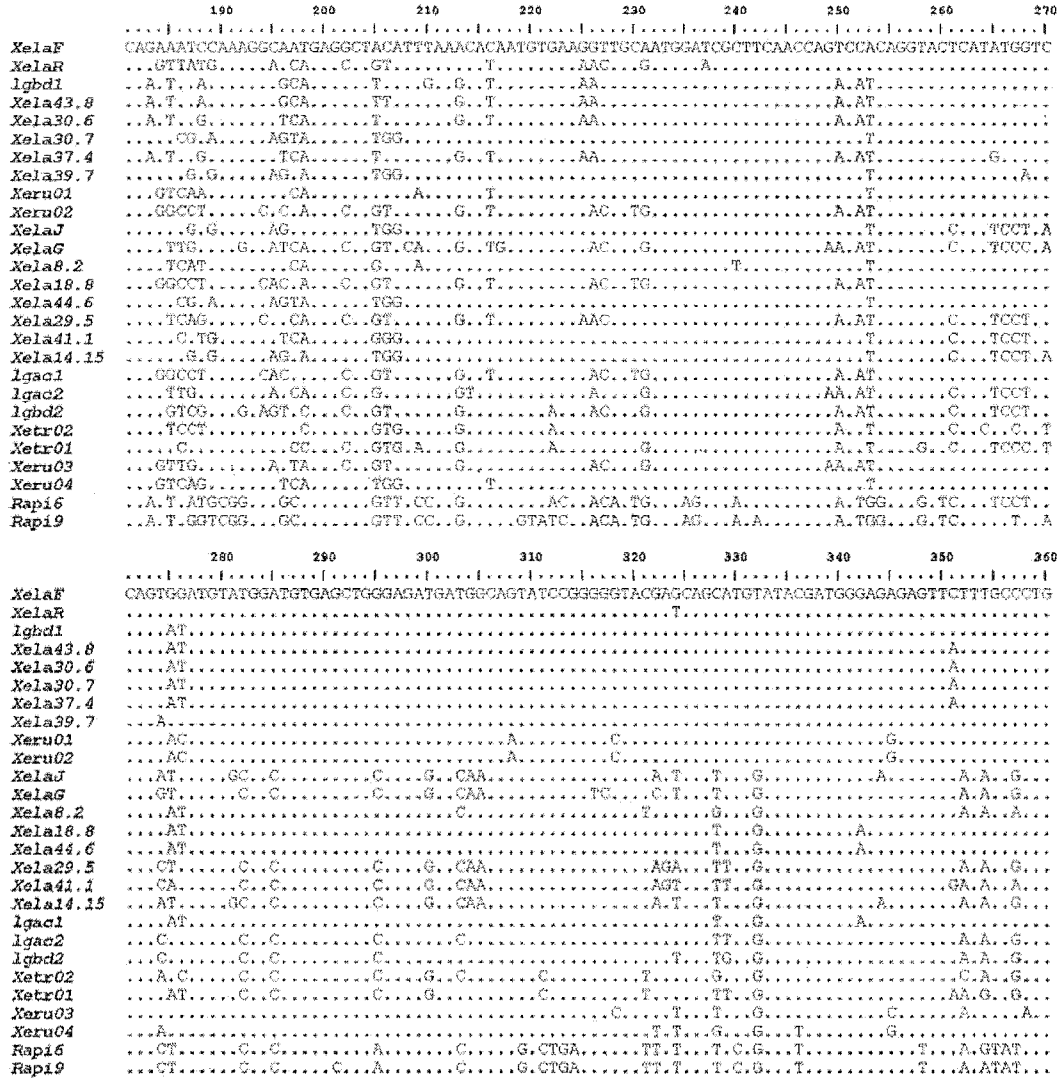
Frog MHC Class Ia Amino Acid Sequence Data



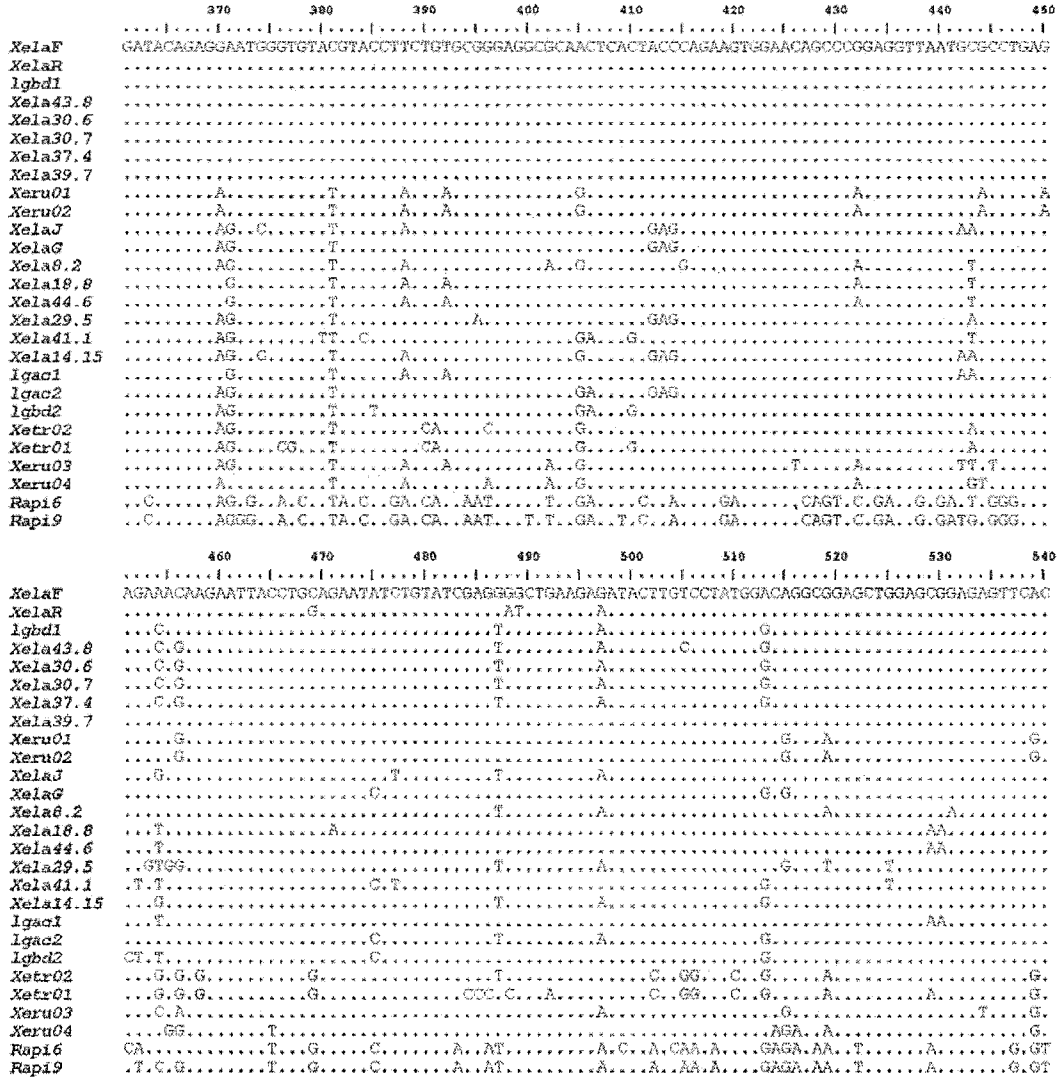
Frog MHC Class Ia Nucleotide Sequence Data



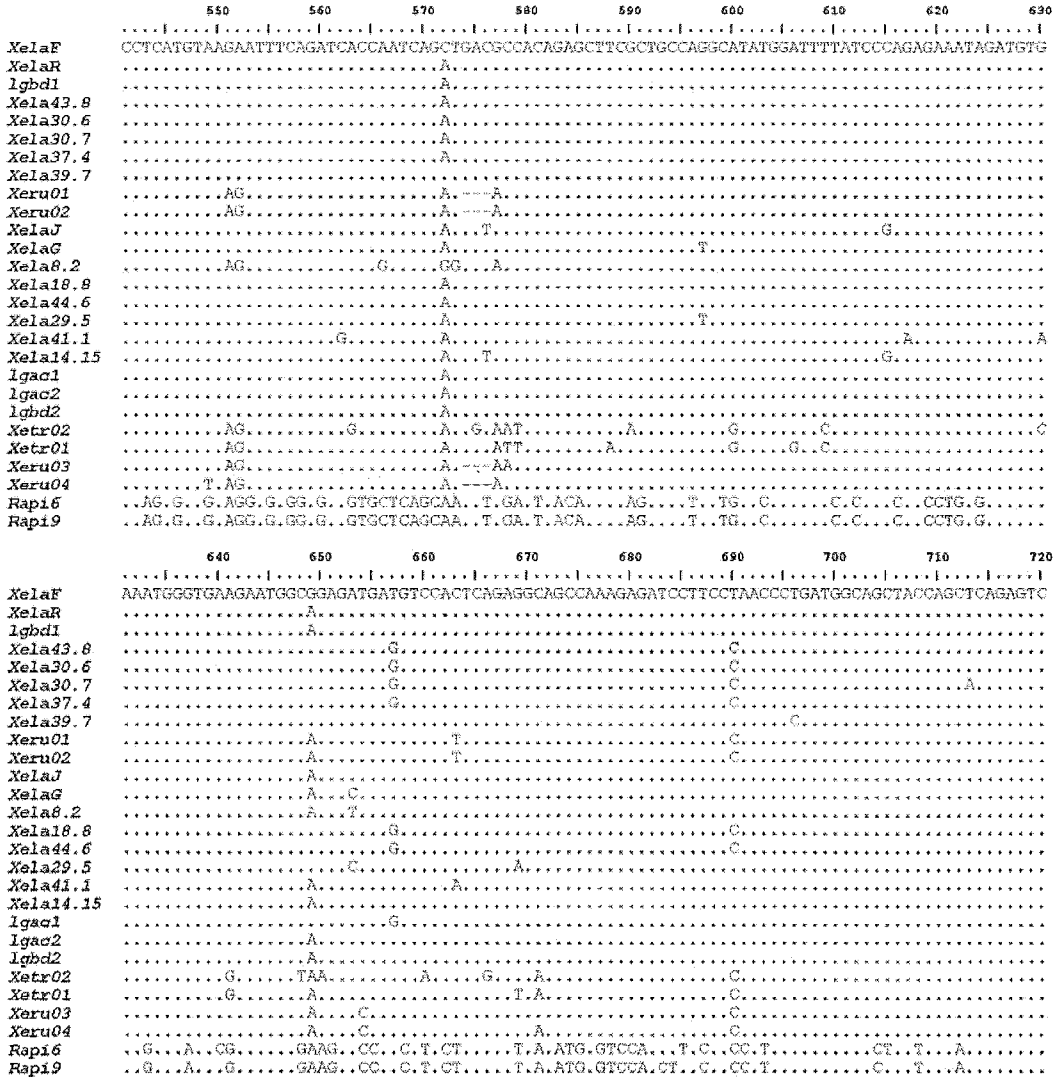
Frog MHC Class Ia Nucleotide Sequence Data



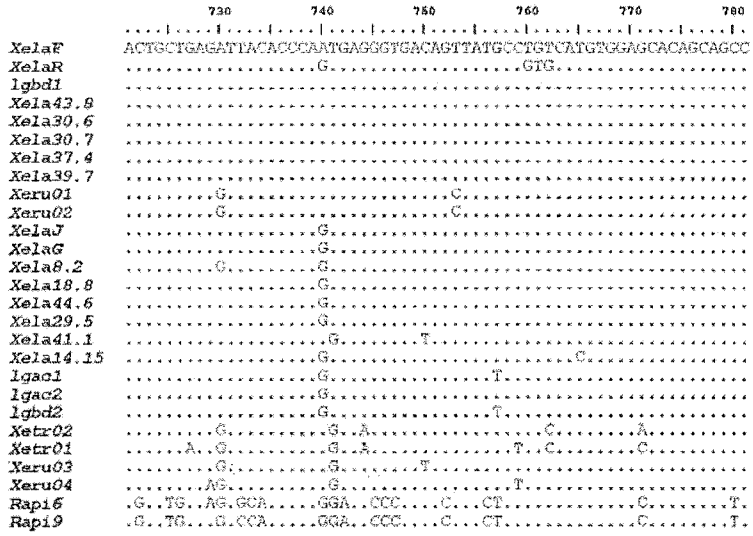
Frog MHC Class Ia Nucleotide Sequence Data



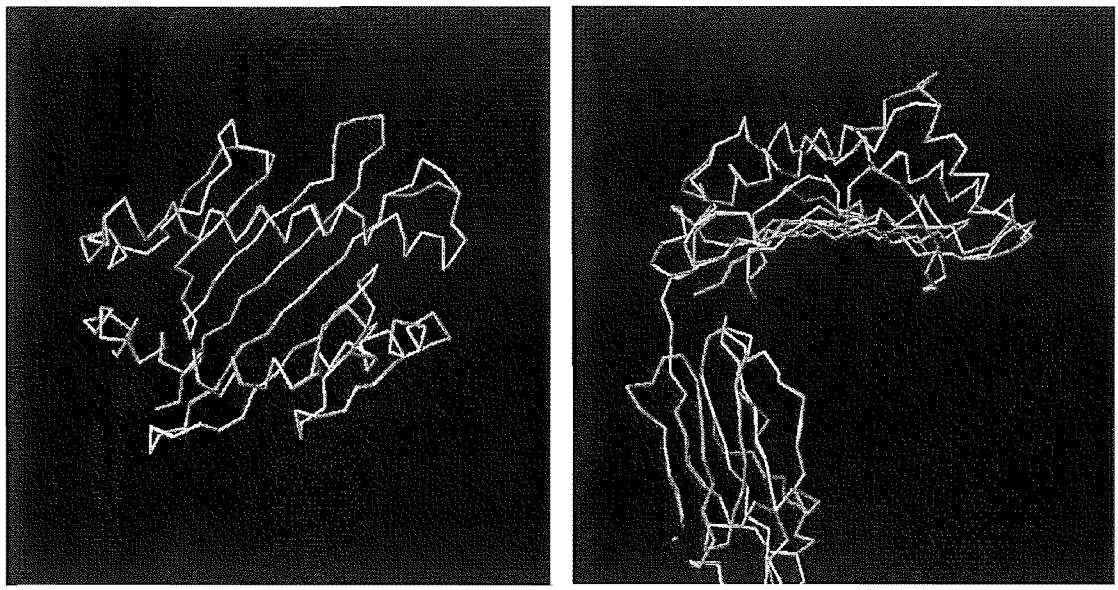
Frog MHC Class Ia Nucleotide Sequence Data



Frog MHC Class Ia Nucleotide Sequence Data



Structural backbone of the *X. laevis* MHC class Ia molecule



A B

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Developmental and Comparative Immunology 29 (2005) 211–227
Developmental & Comparative Immunology
www.elsevier.com/locate/devcompimm

Using models of nucleotide evolution to build phylogenetic trees

David H. Bos[a,*], David Posada[b]

[a]*School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch, New Zealand*
[b]*Departamento de Bioquímica, Genética e Immunología, Facultad de Ciencias, Universidad de Vigo, Vigo 36200, Spain*

Received 10 February 2004; revised 17 June 2004; accepted 31 July 2004
Available online 21 September 2004

Abstract

Molecular phylogenetics and its applications are popular and useful tools for making comparative investigations in genetics; however, estimating phylogenetic trees is not always straightforward. Some phylogenetic estimators use an explicit model of nucleotide evolution to estimate evolutionary parameters such as branch lengths and tree topology. There are many models to choose from, and use of the optimal model for a particular data set is important to avoid a loss of power and accuracy in phylogenetic estimations. Here, we review some molecular evolutionary forces and the parameters included in some common models of evolution used to interpret resulting patterns of molecular variation. We present some statistical methods of selecting a particular model of nucleotide evolution, and provide an empirical example of model selection. Statistical model selection strikes a balance between the bias introduced by some models and the increased variance of parameter estimates that results from using other models.

Keywords: Bayesian phylogenetics; Nucleotide substitution models; Model selection; Akaike information criterion; Likelihood ratio test; Molecular evolution; LMP7

1. Introduction

The use of molecular phylogenetics has become widespread in immunological research because phylogenetic trees are an intuitive way to infer relationships among copies of a gene or among loci of a multigene family. Historically, the primary interest in constructing trees was the pattern of evolutionary relationships itself, or simply the topology of the tree. More recently however, phylogenetic trees are being generated to derive information regarding the processes responsible for the observed pattern of evolutionary relationships, and the tree topology becomes the framework upon which further inference can be drawn. As such, phylogenetics facilitates analysis of gene duplications, rates of evolution, polymorphisms, recombination, divergence of lineages and population

Abbreviations: AIC, Akaike information criterion; BIC, Bayesian information criterion; hLRT, hierarchical likelihood ratio test; *I*, proportion of invariable sites; ln L, log likelihood; LRT, likelihood ratio test; ML, maximum likelihood; MP, maximum parsimony; NJ, neighbor joining; ti, transition; tv, transversion.

* Corresponding author. Address: Department of Forestry and Natural Resources, Purdue University, 715 W. State St, West Lafayette, IN 47907-2061, USA. Tel.: +1 765 494 9779; fax: +1 765 496 9461.
E-mail address: dbos@purdue.edu (D.H. Bos).

0145-305X/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.dci.2004.07.007

Symbols

α	alpha shape parameter	δ	difference of values
Γ	gamma distribution	χ^2	Chi square distribution

demographics [1,2]. Accurate estimates of evolutionary parameters often hinge on the validity of a single phylogenetic reconstruction upon which inference is based. Inaccurate estimation of trees may lead to biased results and erroneous inference of processes or mechanism of evolution.

Several methods of estimating phylogenetic trees are available. Some of the more commonly used methods include neighbor joining (NJ) [3], maximum parsimony (MP) [4] and maximum likelihood (ML) [5]. More recently, new methods that employ a Bayesian statistical approach [6,7] have been successfully implemented, and these methods have quickly generated much interest [2,8]. While several differences exist, one common feature that unites NJ, ML and Bayesian methods is the use of explicit statistical models of nucleotide evolution.

In the context of phylogenetics, a model provides a framework through which the phylogenetic construction method estimates parameters used to find the preferred tree. The model represents the footprint of evolutionary phenomena that has generated the observed sequence data, such as mutation, selection, and genetic drift. The particular model selected for a data set depends on features of the data such as the level of variation and nucleotide frequencies. While it is not our intent to engage in a full review of phylogenetic methods (for reviews see [9–11]), ML, NJ and Bayesian methods generally benefit from their use of models of evolution in terms of flexibility and performance [12,13].

At the outset, the reconstruction of molecular phylogenetic relationships seems a relatively simple exercise. However, the intricacies of DNA sequence evolution and the culmination of molecular forces acting on sequences can make phylogenetic inference a complex matter. The purpose of this paper is to highlight the uses and advantages of nucleotide models in light of the complexities of evolutionary genetics. First we review aspects of DNA sequence evolution such as rates of evolution and changes in those rates through time and along the sequence. We then examine parameters of some models commonly used in phylogenetics that correspond to aspects of sequence evolution and discuss model selection and use. Finally, we present an empirical example of model selection in comparative immunology and use it to demonstrate how results can vary depending on the model being used and argue that appropriate model selection and use is critical to accurate scientific exploration of genetic information.

2. Sequence evolution and phylogenetics

2.1. Substitutions

As more DNA sequences become available, it is apparent that patterns of nucleotide changes used to construct trees are very complex. These complexities arise because of a number of factors contributing to and acting on the primary unit of sequence differences—substitutions. Substitutions can be classified as transitions (ti) or transversions (tv) (Fig. 1). Transitions are substitutions between structurally similar nucleotides (e.g. A↔G, which are both purines), and transversions occur between dissimilar

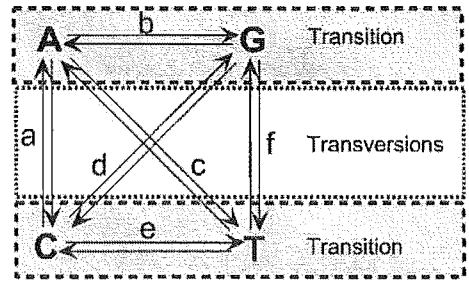


Fig. 1. A substitution matrix representing the possible different rates of evolution for the two possible transitions, and four possible transversions (a–f). In this substitution matrix, substitution parameters are reversible, so that the rate of change from nucleotide i to nucleotide j is the same as rate of change from j to i.

nucleotides (e.g. A↔T; purine to pyrimidine). Transitions are often observed at more than two times the rate of transversions (ti: tv > 2) even though there are twice as many possible transversions for any given nucleotide site. This trend towards more transitions occurs because mutation to a similar nucleotide is more likely to be tolerated than a dissimilar one, and this transition bias can be quite pronounced in some molecules, especially mitochondrial DNA. Frequently, whether or not a substitution is a transversion has implications for altering the protein coded by a DNA sequence.

Substitution rates can vary along a DNA sequence in at least two different ways. First, because of the redundant nature of the genetic code, substitutions are similarly tolerated more or less in various positions within each codon [14]. For instance, the third position of a codon evolves much more quickly than the second position because substitutions at the second position usually change the amino acid encoded by that codon, while similar substitutions at the third position do not. Second, to preserve the function of the protein, its structure must be conserved in important regions; other segments of the protein may be less conserved (for a well known example, see [15]). Thus substitution rates vary in different parts of the DNA sequence correlating to different domains in the protein (i.e. among codons rather than within a codon) and can cause different parts of a gene to support different trees. The variation in substitution rates among different nucleotides in a sequence (rather than in a codon) is referred to as substitution rate heterogeneity or among-site rate variation. In a DNA sequence with among-site rate variation, some nucleotide sites undergo frequent substitutions, while others may change very slowly or not at all [16]. The occurrence of among-site rate variation alters the probabilities of nucleotide substitutions from the often-assumed notion that substitutions are randomly spread along the sequence, and is nearly ubiquitous among DNA sequences [17,18].

2.2. *The molecular clock*

The idea of the molecular clock is based on early observations that the number of amino acid replacements between species or lineages is proportional to the divergence time between them [19]. The empirical observation of a molecular clock was explained by the neutral theory of molecular evolution [20], where such a clock would be expected if most amino acid substitutions were selectively neutral, driven by mutations and random drift. Although the neutral theory has become pervasive in evolutionary genetics, the molecular clock does not always tick regularly [21]. Variation of substitution rates both within a lineage and among lineages makes the existence of a global molecular clock unlikely even though neutral mutations may dominate molecular evolution. Anything that changes the balance between drift and selection can alter the tick-rate of the molecular clock by causing a temporary increase or decrease in the number of substitutions per unit of time, and even neutral evolution can occur in an episodic manner [22,23]. Events such as gene or genome duplications, speciation or changes in the population size can change the dynamic between drift and natural selection, altering the rate of evolution if only for a short period of time.

Many lines of evidence are against a universal molecular clock; however, neutral theory still plays a prominent role in evolutionary genetics. The action of natural selection does not imply that neutral substitutions do not exist, only that they do not always accumulate with clock-like regularity. Violations of the molecular clock are commonly found in highly divergent gene sequences, genes that are the product of gene duplications [24], or genes that have experienced natural selection or changes in structure or function [25]. There are many difficulties associated with using a molecular clock [26], nevertheless, it is often the case that tests of the molecular clock [27] cannot reject clock-like evolution for closely related gene sequences. This could indicate that molecular evolution is clock-like for periods of evolutionary time, or that methods may lack statistical power to reject a molecular clock in some cases. Even when clock-like evolution is plausible, precise estimation of dates can still be difficult to obtain because of different assumptions and sources of uncertainty [28]. Also, methods are available that relax the assumption of a strict molecular clock and allow one to estimate evolutionary dates in lineages that have different rates [29–31].

Many evolutionary processes create irregular patterns of nucleotide substitution and the detection

and characterization of these irregularities has led to a better understanding of DNA sequence evolution. In turn, our understanding of molecular evolutionary patterns has allowed us to develop statistical models used to represent the irregularities of DNA sequence evolution. For instance, through the use of these models, researchers are able to overcome common phylogenetic scenarios that are positively misleading for methods that do not use statistical models such as MP [32–34]. Although models are ultimately major simplifications, summarizing many evolutionary forces and events, appropriately incorporating these models generally leads to improvement of genetic distance and phylogenetic analysis [11].

3. Models of nucleotide substitution

3.1. Phylogenetic estimators

Neighbor Joining, ML and Bayesian methods all rely on explicit statistical models of evolution to reconstruct evolutionary trees. The Neighbor Joining algorithm is different from ML and Bayesian methods because it uses the model to calculate pairwise genetic distances between sequences, and reconstructs a topology based on those distances. Maximum likelihood and Bayesian methods use the sequence data directly to reconstruct a tree, thereby utilizing information in specific nucleotide differences instead of summarizing changes with a genetic distance. Due to these differences, ML offers noteworthy statistical properties in comparison with genetic distance-based methods, but is much more computationally intensive [32,35,36]. While NJ and ML methods are well understood and their uses are common in the literature, Bayesian methods are relatively new.

The Bayesian method is related to ML method because they both utilize the likelihood function. However, when using Bayesian statistics to reconstruct a phylogeny, the preferred outcome is the one that maximizes the posterior probability, which is determined by the prior distribution and the likelihood of that tree. The prior distribution for trees, models and parameters can be specified to be generally uninformative to avoid bias, or it can reflect prior knowledge from other sources. Whereas other methods produce a single best estimate of evolutionary relationships and ignore uncertainty of the final outcome, Bayesian methods produce a set of trees of which the one with the highest posterior probability is accepted as the preferred tree. Bayesian methods are generally faster than ML methods, and also offer the advantage of automatically incorporating an estimate of phylogenetic uncertainty [6]. While many aspects of Bayesian phylogenetic estimation have yet to be refined and explored, these methods offer the same benefits from employing statistical models as ML and NJ [6,7]. These benefits include the flexibility to incorporate a wide range of models, easy hypothesis testing, and improvements on estimates of numbers of substitutions, efficiency and robustness [37].

3.2. Model parameters

Statistical models of nucleotide change represent aspects of the pattern of variation that results from the process of evolution. Models vary in complexity according to the number of parameters used to represent evolutionary change. While simple models summarize nucleotide substitutions with one or two parameters, the most general models can involve more than 60 parameters (e.g. codon models that are introduced below). Model parameters can reflect differences in nucleotide frequencies, substitution rate (such as transition bias) and among-site rate variation. The substitution matrix of a model represents different rates of evolution between certain pairs of nucleotides, and the gamma distribution models among-site rate variation. In other words, the substitution matrix determines the substitution rate between specific nucleotide pairs (e.g. A↔G), and the gamma distribution determines the overall substitution rate at a nucleotide site. Combining different parameters has resulted in a large number of models, but many of them share several parameters (Table 1).

The JC69 model [38] considers all possible nucleotide substitutions to have an equal probability, and is the simplest available model (Table 1). Felsenstein [5] suggested a model in which probabilities of nucleotide changes were determined by the equilibrium nucleotide frequencies. Kimura [39] proposed a model that utilizes a relatively simple substitution matrix that allows for two different rates: one for transitions and the other for transversions.

Table 1
Some commonly used nucleotide models and summary of parameters

Model	Parameters			
	Number of parameters	Nucleotide frequencies	Substitution rate in Fig. 1	Reference
JC69	1	Not included	$a=b=c=d=e=f$	[38]
F81	4	π_A, π_C, π_G, π_T	Not included	[5]
K80	2	Not included	$a=c=d=f$, $b=e$	[39]
K81	3	Not included	$a=f$, $b=e$, $c=d$	[40]
HKY85	6	π_A, π_C, π_G, π_T	$a=c=d=f$, $b=e$	[42]
SYM	6	Not included	a, b, c, d, e, f	[108]
TrN	7	π_A, π_C, π_G, π_T	$a=c=d=f$, b, e	[44]
GTR	10	π_A, π_C, π_G, π_T	a, b, c, d, e, f	[109]

Parameters of these models can include four different base frequencies and up to six substitution rates. Flexibility of models is such that invariable sites and/or a gamma distribution can simply be added to incorporate rate variation.

Kimura [40] and others [41,108] have also formulated models that incorporate more than two rates in the substitution matrix, thus enabling models to account for different rates of change between all of the possible nucleotide pairs. In an effort to make models more representative of empirical observations, Hasegawa et al. [42], Felsenstein [43], Tamura and Nei [44] and Rodriguez et al. [109] each created models which incorporate multiple aspects of sequence evolution (Table 1). These models combine parameters for differences in substitution rates and differences in nucleotide frequency.

Among-site rate variation can also be incorporated into models of nucleotide evolution. The simplest way to statistically represent among-site rate variation is to divide sites into two classes: those that vary and those that are invariable. To better account for wide rate differences among the variable sites, several methods have been used [44,45], but the most successful involves the use of a gamma distribution [18,46]. The gamma distribution can be approximated with as little as four categories [47], and the statistical representation of rate variation is independent of substitution models like those described above and can simply be added to any pre-existing model (for example, we can specify a JC69+Γ model).

Under the gamma distribution, there is a continuum of probabilities of change for nucleotides, ranging from low to high. The numbers of nucleotide sites with the various rates of substitutions determines the shape of the gamma distribution that is summarized by the shape parameter (α). When most of the nucleotides are invariable or have very slow rates, then the shape of the distribution is skewed to the right (Fig. 2). Under this scenario there are a few nucleotides with high rates and the shape parameter would be small ($\alpha<1$), indicating a high level of rate variation, i.e. not all nucleotides evolve at a similar rate. As a result, most of the variation in the data set comes from relatively few nucleotide sites that are evolving very rapidly (substitutional 'hotspots').

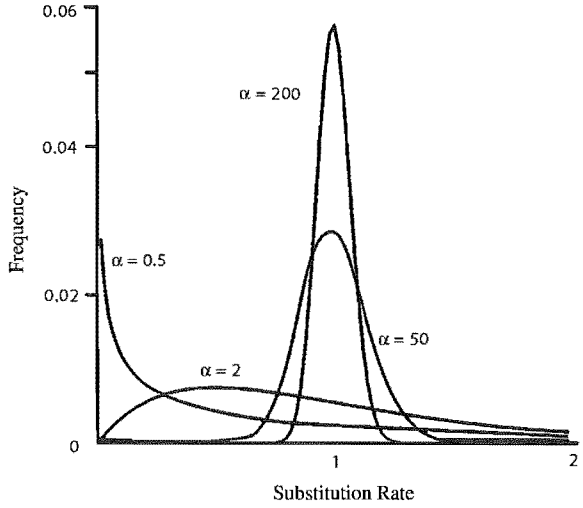


Fig. 2. Gamma distributions calculated using different shape parameters (α). The number of nucleotides in a sequence evolving at a particular rate determines the shape parameter. When a sequence contains mostly invariable nucleotide sites and variation is concentrated at a few rapidly evolving nucleotide sites, the shape parameter is small (<1). As the proportion of variable nucleotide sites increases, the shape parameter becomes larger, indicating that more sites evolve at a moderate rate and fewer sites have extremely high or low rates.

Large shape parameters ($\alpha > 20$) indicate a more bell-shaped distribution with most sites having intermediate rates of evolution with few nucleotides evolving at very high or low rates (Fig. 2). As the shape parameter becomes larger, more nucleotide sites have a more similar rate of evolution and among-site rate variation becomes increasingly inconsequential [13].

The above-mentioned model parameters all work at the individual nucleotide level, and therefore treat each nucleotide as an independent unit. However, for protein coding DNA sequence this is not the case. Whether or not a substitution changes an amino acid depends on the other nucleotides in that codon when the substitution occurs, thus individual nucleotide sites in protein coding sequence are not independent. To accommodate this, nucleotide models that treat a codon triplet as an independent unit have been formulated to more accurately model coding DNA [48–50]. Variations of these models provide parameters to account for transition bias, codon frequency, rate variation among codon positions, and different rates for non-synonymous substitutions [51]. Codon models can become very complex by parameterizing each codon frequency, but these models can also approximate codon frequencies with fewer parameters. Unfortunately, these models are generally not implemented for use in reconstructing phylogenetic trees except when using some Bayesian methods [7]. Instead, codon models have been typically used to estimate substitution rates and detect levels of natural selection acting on a protein.

3.3. Effects of models

The performance of a model-based phylogenetic method may depend on the fit of the model to the data [10]. Similarly, the efficiency of distance-based methods is dependant on the accuracy of model-based estimates of genetic distance [11]. For sets of sequences that are long with low levels of polymorphism, the model may have little effect on the outcome of analysis. However, when working with more divergent sequences, the use of one model over another can alter the results of analysis, and even lead to strong support for the wrong tree topology [52], a fact that underscores the importance of using the best-fit model for a particular data set. Due to the wide diversity in size, variation and rates of evolution among different data sets, there is no single best-fit model suited for use in any data set. Use of inadequate, overly simplistic models selected without statistical validation often leads to biased estimation of evolutionary genetic parameters [12,33,37,53,54].

The model parameter with one of the strongest influences on genetic distance and phylogenetic estimation is among-site rate variation. Rate variation among sites is particularly problematic and misleading when substitution rates also vary among branches in the tree (e.g. non-clock-like evolution) [32]. When both types of variation are present, use of the best fit model seems to be essential to obtain the correct tree topology [16,55]. Except in cases with strong rate variation among both sites and lineages, tree topology estimation is relatively robust to violations of model assumptions [36,56]. Unfortunately the same robustness does not extend to estimation of parameters such as substitution rates, branch lengths and genetic distance. Failing to include rate heterogeneity among sites results in underestimation of the number of substitutions at highly mutable sites [16]. Consequently, branch lengths are underestimated, and this effect is much more prominent in longer branches than shorter ones [54]. This is likely to be due to the fact that phylogenetic estimators give greater weight to highly variable sites in a sequence [47].

Simplifying the assumptions of a model by failing to include a factor for transition bias can also adversely alter the outcome of analysis. A transition bias is found universally among DNA sequences [57] and inclusion of this parameter is essential for accurate estimates of genetic distance for NJ analysis [58,59]. Similarly, failure to incorporate transition bias will result in underestimation of branch lengths in ML phylogeny estimation [60]. Aside from the inherent problems of branch length and genetic distance underestimation, these factors can alter the tree topology and lead to erroneous conclusions regarding the dates of lineage splitting [44]. There is also an interplay between transition bias and among-site rate variation, so that the level of among-site rate variation is underestimated (overestimation of α) using models that exclude a transition bias [60].

One of the major advantages of using models is the ability to more accurately estimate the actual number of substitutions that have occurred in a set of sequences. This allows researchers to include

sequences of high variability because homoplasy in the form of superimposed substitutions can be accounted for with the use of models. The alternative way of dealing with sites or sequences which are suspected of saturation of substitutions, is simply to eliminate then from consideration. While this does effectively eliminate the influence of homoplasy at those sites, any information that can be gleaned from those sites is also lost and the size of the sample is decreased, exposing the analysis to the increasing effects of sampling error or bias.

While potential problems with simple models are documented, some also dispute the utility of more general models [61]. Some criticisms of very complex models point out that these models have greater difficulty distinguishing between tree topologies because of smaller differences in likelihood scores, and that as more model parameters are added, more error is associated with each parameter estimate. These properties of complex models are general statistical phenomena and are not limited to phylogenetic analysis; however, while these points are valid, they arise because of random rather than systematic error. As a result these problems can be mediated rather than aggravated by addition of data [13]. The amount of data required for consistent phylogenetic analysis depends on the shape of the tree, numbers of taxa and levels of diversity. If the tree shape is not symmetric and branch lengths are very long, then analysis of data with less than 500 nucleotides will generally not be reliable, especially for more general models [33,62]. Consistency and reliability of phylogenetic inference is expected to increase by analyzing longer sequences and additional taxonomic sampling.

The potential bias introduced through using a particular model also has an effect upon the level of support given to a tree topology with techniques like bootstrapping [63]. The most widely accepted interpretation of the bootstrap is that it is the level of support for a particular node of a tree that the data provides [64]. As such, it represents whether the same topology might be recovered if more data are collected, rather than if the relationship is correct. However one interprets the bootstrap values, the accuracy and precision of the bootstrap values depends on the fit of model [65]. For instance, if a phylogenetic method or a model is used that has systematic bias, then the bootstrap will also reflect that bias [13]. Consequently, bootstrap values used in such a case will be artificially high and reflect strong support for incorrect branching patterns.

Bayesian methods estimate a level of phylogenetic support that is seen as an intuitive measure of uncertainty regarding each tree topology. Less work has been done to evaluate Bayesian measures of support and the relationship of model specificities and levels of support [66]. However, some research shows that Bayesian measures of support are good estimates of phylogenetic accuracy [67], but others conclude that these values are overestimates of the true level of uncertainty [68,69]. Regardless of the procedure used to measure phylogenetic support, caution interpreting results is warranted and use of a statistically rigorous method of selecting a model is recommended.

Although conflicting examples of model complexity and phylogenetic accuracy can be found [65,70], one trend that has emerged is that because of the increase in variance, very short sequences (which are statistically equated with small sample sizes) often do not support the use of the same level of model complexity as longer sequences. Even though the underlying evolution of short sequences may be just as complex as longer sequences, the larger variance inherent with generalized models and small sample sizes makes these types of data more prone to the effects of over-parameterization [71]. While the relationship of model parameters and performance of Bayesian, ML and NJ tree estimation is not always straightforward, a trade-off between the bias of simple models and the increased variance of more general models is generally observed [12]. Consideration of models should take into account the size of the data set, level of divergence, amount of differences in substitutions between different nucleotides, and constancy of rate of evolution both in time and along the sequences. Use of objective criteria to select models will help avoid problems associated with model over-fitting by ensuring that models are not excessively complex and avoid phylogenetic bias by selecting more realistic models [72].

Models that can be implemented in popular phylogenetics programs such as PAUP* [73], PHYLIP [43], MEGA2 [74] and MRBAYES [7,75] are useful approximations of DNA sequence evolution. Use of one particular model versus another often changes the outcome of analysis, and the choice of

models can be more important than the method of phylogenetic reconstruction. Given that the model plays a great role in the results of analysis, it seems that the choice of one model over another should be justified in some way. Unfortunately, it is still commonplace for models to be used indiscriminately and without justification. The question then becomes, which model is appropriate for a particular data set and how can that model be justified?

3.4. Model selection and use

To minimize adverse effects of model over-fitting and model under-fitting, the ideal use of models is to incorporate as much model complexity as needed and no more. Fortunately, methods for selecting the most appropriate model for a particular data set have been proposed. These methods provide a rigorous statistical framework in which to select and justify the best fit model. With the goal of finding the simplest model that accurately approximates sequence evolution, Rzhetsky and Nei [76] developed statistics for selecting models. These tests are independent of evolutionary time and do not require an a priori phylogeny on which to base inference. While this method is computationally efficient, its application is model-specific and restricted to a limited subset of the available models.

Another method is to use the likelihood ratio test (LRT) to compare models [10]. The LRT statistic is calculated by obtaining the likelihood scores of a null model (L_0) and an alternative model (L_1). The two scores are then compared by taking twice the difference in the logarithm of the likelihoods to obtain the statistic $[\delta = 2(\ln L_1 - \ln L_0)]$. Use of the LRT in phylogenetics is commonplace for hypothesis testing and the distributions and performance of the test have been investigated [77,78]. When the models compared are nested (one is a special case of the other), the Chi-square distribution (χ^2) is a good approximation of the null distribution of the LRT statistic (df = the difference in the number of free parameters in the two models). In some special cases, fixing one of the parameters of the more parameter-rich model at either boundary (0 or ∞) reduces the model to the simpler null model, and a mixed distribution is used [79].

The LRT can be performed on any of the available models, but it requires an a priori input phylogeny to estimate the likelihood of the models [80]. It is also easy to test several models against each other in a series of LRTs that can be performed in a hierarchical fashion (Fig. 3). The likelihood scores of the two

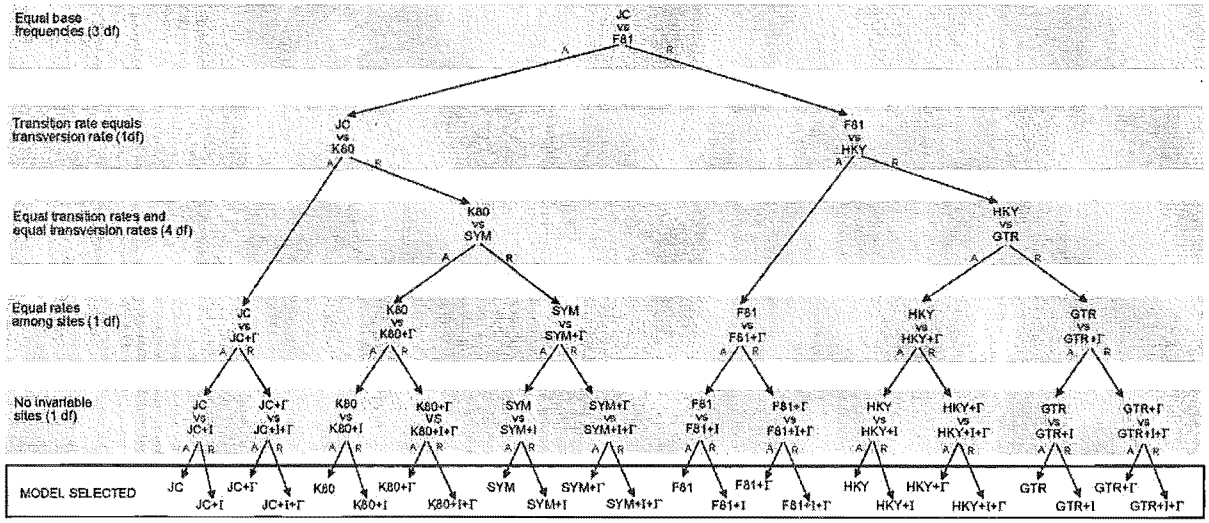


Fig. 3. A 'decision tree' of a hierarchical likelihood ratio test. Hypotheses tested are indicated on the left, and this schematic begins with the simplest model and progresses to more complex models in a stepwise manner. The pathway chosen depends on acceptance or rejection of LRT scores, based on a chi squared distribution ($P < 0.01$). In order to preserve the clarity of the figure, not all available models are shown. Models depicted here are: JC [38]; F81 [5]; K80 [39]; HKY [42]; SYM [108]; GTR [109]. *I*, proportion of invariable sites; *Γ*, gamma distribution of rates among sites.

Table 2
Several models are compared successively to determine the best fitting model for a data set, starting with the simplest model and increasing complexity

Null model	Alternative model	Parameter tested	LRT (δ)	P-value
JC69	F81	Equal base frequencies	4.680	0.196
JC69	K80	ti=tv	90.022	0.000
K80	SYM	Equal ti and tv rates	50.340	0.000
SYM	SYM+Γ	Equal rates among sites	314.512	0.000
SYM+Γ	SYM+Γ+I	No invariable sites	14.866	0.000

The parameter being tested is assumed by the current null model but not the alternative model. The null model is rejected when the P-value of the LRT is <0.01 using a χ^2 or mixed χ^2 distribution.

models are compared using the LRT test statistic, δ, and significance of the LRT statistic is determined. The better fitting model is retained, it becomes the null model, and the process is iterated with successively more general models of evolution until the addition of further complexity in the alternative model does not create a significantly better fit to the data (Fig. 3 and Table 2). The LRT may be appealing, but the significance of LRTs are easily calculated only for nested models and the a priori distribution of significance for non-nested comparisons is not well established. Performance tests of the LRT also show that this criterion is good at recovering the model used to simulate the sequence data [80], although we should keep in mind that in reality the true model of nucleotide substitution is unknown, and it is much more complex than any candidate model that we can select.

Another way of selecting the most appropriate model for a data set is to use the Akaike information criterion (AIC) [81], which can be thought of as the amount of information lost when a particular model is used to approximate reality. The AIC implements best-fit model selection by calculating the likelihood of proposed models, and imposing a penalty based on the number of model parameters. Parameter-rich models incur a larger penalty than more simple models so that fitting an excessively complex model is not likely. The best fitting model is the one with the smallest AIC value, ($\mathrm{AIC} = -2 \ln L_i + 2N_i$), where L_i is the likelihood for model i and N_i is the number of free parameters in model i. Although the use of LRTs is much more extended in phylogenetics than the use of the AIC, the latter offers important advantages [71]. The AIC is able to compare non-nested models and simultaneously compares all candidate models, rather than performing sequential pair-wise comparisons; the AIC also has a simple adjustment that more heavily penalizes complex models for data comprised of small samples (i.e. short sequences). The AIC also allows for model selection uncertainty and model averaging. In addition, the AIC recognizes that the true model is not among the set of candidate models so it tries to find the candidate model that best 'approximates' the true unknown model of molecular evolution given the amount of information in the data. The objective of model selection is to find the model that will allow one to most accurately estimate unknown phylogenetic parameters while avoiding bias and excessive variance. The model that is best suited to that end will not be an exact representation of cumulative evolutionary processes, but a useful approximation that is appropriate for the level of polymorphism and size of the data set.

Bayesian statistics have also been adapted for use in phylogenetic model selection. Bayes Factors make pairwise model comparisons and are therefore analogous to the LRT procedure [82,83]. Alternatively, the Bayesian information criterion (BIC) can be used [84]. This method more easily enables comparisons of multiple models and is easy to calculate. The posterior probabilities of Bayesian statistics are already used to discriminate between phylogenetic trees and these measures can also be used to choose among multiple models [85]. Like the AIC, Bayesian methods allow estimation of model uncertainty and allow estimation of a phylogeny using a set of candidate models in a model averaging procedure. An important distinction of Bayesian statistics is that calculation of likelihoods proceeds

differently, so that likelihood values compared using Bayesian methods are different from those used in AIC or LRT comparisons.

The above techniques compare model fitness relative to other candidate models, but measuring overall adequacy of a model can also be done. To do this, Navidi et al. [86] and Goldman [87] describe a test that compares a model with an unconstrained model and the appropriate distribution to test significance. Also, Bayesian methods have recently been adapted to examine the adequacy of models [88]. While the unconstrained model is very complex, it is worth noting that when comparing any two models, only aspects in which the models differ are tested. Any aspects models have in common or aspects that are not included in either model remain untested. The outcome of general adequacy tests may find that the selected model is not a complete representation of the data. This is usually thought to be the result of the stringency of the test, instead of gross misrepresentation of the data by the model. Rather, this outcome simply means that the model does not perfectly describe all of the underlying processes of molecular evolution, as would be expected.

The impact of models on phylogenetic analysis is very significant, strongly affecting branch lengths and often topology as well. The use of any particular model is not wrong per se, but we advocate statistical, objective selection among available candidate models to maximize the use of available models for each data set. Unfortunately the model used for analysis is often not justified or even reported in the literature despite its influence on the outcome. However, easy-to-use computer programs that implement rigorous statistical selection of models are available [89]. In the following we demonstrate their use and show how model selection determines the outcome of phylogenetic analysis.

4. Empirical example

4.1. Data

To illustrate aspects of model selection, we reconstructed the phylogenetic relationships of nine taxa using DNA sequences of the LMP7 gene downloaded from the Genbank database (accession numbers: human, *Homo sapiens* BC001114 (the human LMP7 is also termed PSMB8 or RING10); mouse, *Mus musculus* U22032; frog, *Xenopus laevis* D44540; salmon, *Salmo salar* AF184938; zebrafish, *Danio rerio* AF032390; medaka, *Oryzias latipes* D89725; pufferfish, *Fugu rubripes* AJ271723; nurse shark, *Ginglymostoma cirratum* D64057; horn shark, *Heterodontus francisci* AF363583). Copies of the gene are from a variety of vertebrates from which full length cDNA was obtained, and the estimated phylogenetic relationships could be used in the framework of studying multigene family evolution, estimating substitution rates, or establishing homology of gene copies. The leader peptide was excluded from analysis, leaving only the coding sequence from the mature protein. The sequences were aligned using Clustal W [90] and alignments were inspected to ensure that the integrity of the coding frame was preserved. The best-fit model for these data was selected using the LRT and AIC after calculating likelihood scores of 24 models using PAUP*4.0 [73].

4.2. Methods

The best fitting model for these data was evaluated according to a hierarchical LRT. The AIC method of model selection was also used to find the best-fit model by calculating the likelihood and subsequently the AIC score of all models. Phylogenetic trees were also calculated using 24 models of evolution selected to represent a variety of statistical complexity. These models have an arbitrary relationship to the data, and the resulting trees can be compared to those obtained using models selected using rigorous statistical criteria. Here, we use the ML method of phylogenetic construction as implemented in PAUP* [73] because it is known to be robust to violations of model assumptions and because the statistics of ML estimation are well understood [10,36,56]. We calculated these scores manually to demonstrate the method, but the program MODELTEST [89] provides the appropriate command block for PAUP* to automatically calculate the likelihood scores for 56 models, which can then be automatically compared using the LRT and AIC in the MODELTEST program.

4.3. Results

The model selected by both the LRT and AIC is the SYM model with both invariable sites and a gamma distribution of among-site rate variation (SYM+Γ+I; see Tables 2 and 3) [47,108]. This model includes a substitution matrix allowing for six different rates of substitutions: one for each type of reversible nucleotide change. There is no significant heterogeneity of nucleotide frequencies accounted for in the model, but the model makes provisions for considerable rate variation along the gene sequence (see Table 3). The invariable sites of the sequence alignment are accounted for in the model and the gamma distribution represents rate heterogeneity only among variable sites. As the distribution of the gamma shape parameter is skewed towards the right, most of the variable nucleotides evolve fairly slowly, with a few sites evolving more rapidly. Models with few parameters commonly used to reconstruct phylogenetic relationships were rejected by both selecting criteria in favor of more general models (Table 2).

A test of the overall adequacy of the preferred model against the unconstrained model [87] indicates a sufficient level of support for the SYM+Γ+I model. The test statistic of the difference in likelihoods between the unconstrained and SYM+Γ+I models was 1091.546. Monte Carlo simulations under the null (SYM+Γ+I) model hypothesis were done to determine the null distribution of differences in likelihood between the unconstrained and SYM+Γ+I models. This distribution ranged from 946.723 to 1196.620 with a mean value of 1069.492. The test statistic falls well within the 95th percentile of the distribution, indicating that the null hypothesis (SYM+Γ+I) cannot be rejected against the unconstrained model under these criteria. The best-fit model selected above was therefore used to reconstruct the phylogenetic relationships among these taxa, and the result indicates a topology consistent with generally accepted relationships (Fig. 4).

When the evolutionary relationships among these genes were estimated using other models, three different tree topologies emerged (models and likelihood scores found in Table 4). Many simple models rejected by statistical model selection criteria preferred a tree in which the frog and shark share a most recent common ancestor, and this clade is a sister group to a clade in which mammals and teleost fish

Table 3
Molecular evolutionary parameter values of best-fit model, SYM+Γ+I, selected under the LRT and AIC criteria

Parameter	Value
Substitution matrix	
A: C	1.49
A: G	2.12
A: T	1.53
C: G	0.62
C: T	4.24
G: T	1.00
Base frequencies	
A	0.25
C	0.25
G	0.25
T	0.25
Proportion invariable sites	0.411
Gamma shape parameter	2.827

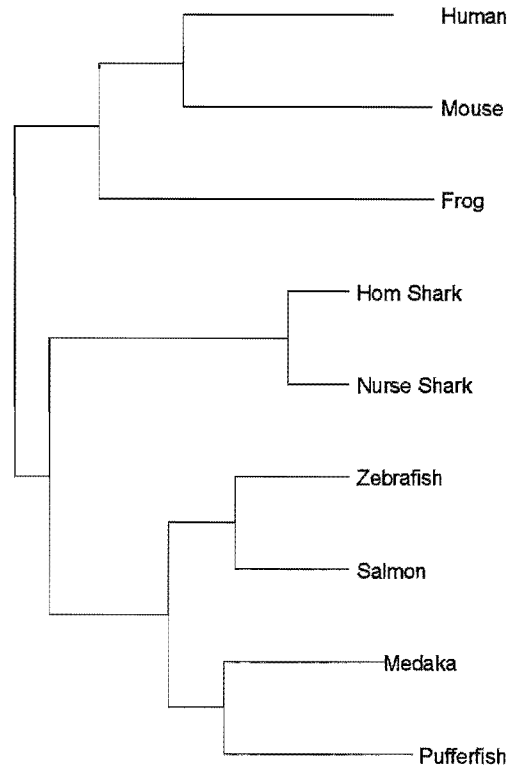


Fig. 4. Unrooted phylogenetic tree generated using the maximum likelihood optimality criterion and the preferred model of nucleotide evolution (SYM+Γ+I) selected by the hierarchical likelihood test and the AIC criterion.

Table 4
Log likelihood scores ($-\ln L$) of models calculated using the single NJ tree topology used in the hLRT

Model	$-\ln L$	δAIC	Topology
JC	3813.934	455.742	Fig. 5a
JC+I	3652.954	135.782	Fig. 5a
JC+Γ	3659.403	148.680	Fig. 5a
JC+I+Γ	3650.366	132.606	Fig. 5a
F81	3811.594	455.062	Fig. 5a
F81+I	3651.362	136.598	Fig. 5a
F81+Γ	3657.217	148.308	Fig. 5a
F81+I+Γ	3648.494	132.862	Fig. 5a
K80	3768.922	367.718	Fig. 5b
K80+1	3602.037	35.948	Fig. 4
K80+Γ	3606.883	45.640	Fig. 5a
K80+I+Γ	3598.376	30.626	Fig. 5b
HKY	3766.357	368.588	Fig. 5b
HKY+I	3601.262	40.398	Fig. 4
HKY+Γ	3605.544	48.962	Fig. 5b
HKY+I+Γ	3597.331	34.536	Fig. 4
SYM	3743.752	325.378	Fig. 4
SYM+I	3583.904	7.682	Fig. 4
SYM+Γ	3586.496	12.866	Fig. 5b
SYM+I+Γ	3579.063	Best	Fig. 4
GTR	3737.487	318.848	Fig. 5b
GTR+I	3582.327	10.528	Fig. 4
GTR+Γ	3583.892	13.658	Fig. 5b
GTR+I+Γ	3576.966	1.806	Fig. 4

Significance of likelihood comparison summarized in Table 2. Topology reconstructed under each of 24 models representing various levels of complexity. See the caption of Fig. 3 for model references.

share a most recent common ancestor (Fig. 5a). Eight of the nine models that reproduce this topology share a common feature: they do not have a substitution matrix specifying different rates for substitutions between different nucleotide pairs. Seven other models reconstructed a tree in which frog and mammals formed a tetrapod clade and fishes formed a monophyletic group. However, in this tree, the pufferfish and medaka, generally considered derived fishes, are found at the root of the teleost clade, displacing the more primitive salmon and zebrafish (Fig. 5b). In total, eight models preferred the 'correct' topology; however no clear pattern of which models reconstruct the 'correct' tree exists for these data (Table 4). For example, not all models that include a parameter for among-site rate variation result in the 'correct' tree, and some models that are more complex than the best-fit model found the 'correct' tree and some reconstructed another topology. The lack of a clear pattern in progression of model parameters and tree structure illustrates that it is often impossible to tell a priori which models will find the same tree as the best-fit model, a fact that underscores the importance of finding the best-fit model.

4.4. Discussion

It is difficult to fully assess why some models reconstruct a topology inconsistent with generally accepted taxon relationships in this example, and multiple factors of sequence evolution are often the cause. In this case, the level of diversity may contribute to misleading results. A very high level of diversity means that many potential substitutions may be unaccounted for using simple models that consistently underestimate the number of substitutions for distantly related species [60]. Multiple substitutions at given sites may provide conflicting evidence for various relationships, weakening support for a clade or overall branching pattern. This lack of consistent support renders trees with different topologies statistically indistinguishable. We tested the statistical difference among trees using the Shimodaira–Hasegawa test [91], and found no significant difference between all three topologies (Fig. 5a, $P=0.305$; Fig. 5b, $P=0.572$). Since some more complicated models also fail to reconstruct the widely accepted 'true' phylogeny it is likely that other factors play a role in misleading phylogenetic analysis. For these data, other factors such as different rates of evolution in part of the tree may also decrease the usefulness of models.

Use of simplistic models in evolutionary genetics can be misleading if the sequences do not evolve according to a molecular clock [92]. We used a simple LRT to test whether or not these sequences evolve according to a molecular clock [5] to see if this may be a misleading factor for simpler models. The LRT statistic was 137 ($P<0.0001$; df=7) indicating that the model enforcing a strict molecular clock was a much worse fit to these data. These results corroborate those of Takezaki et al. [93], who found variable substitution rates among lineages of proteasome components. Most statistical models of nucleotide evolution are 'stationary,' in that model parameters are constant across the entire tree; however, non-stationary models have been formulated that allow

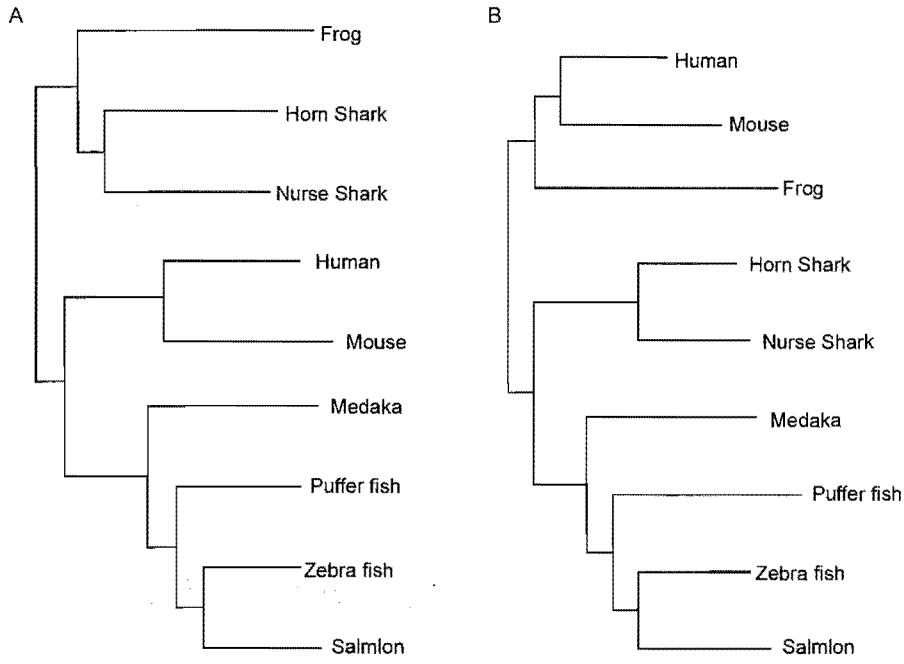


Fig. 5. Unrooted phylogenetic trees generated using the maximum likelihood optimality criterion. Twenty-four different models of nucleotide evolution (Table 4) were systematically selected to represent a range of models with differing levels of complexity, but arbitrarily selected with regard to how well they fit the data. These models were then applied to the data, and several of these models supported trees with topologies that differed from that reconstructed using the optimal model.

parameters to change with time [94,95]. Use of these models generally improves the fit to the data and performance of the method, but greatly increased model complexity. Here, the overall rate of evolution is different among branches of the tree, therefore this and other simplifying assumptions may affect model fitness and utility.

Other factors may be involved in the failure of some models, as Whelan et al. [96] indicate that positive or negative selection may be an unaccounted for dynamic that affects phylogenetic reconstructions. For instance, selective pressures can result in convergent evolution, causing divergent taxa to appear closely related. The test of model adequacy indicates support for the SYM + Γ + I model, but both models in that comparison make no provisions for natural selection and they both assume that data at each site is independent and identically distributed. Therefore, neither of these aspects of sequence evolution is evaluated in this comparison. Due to the coding nature of these sequences it is likely that both natural selection and non-independence of nucleotide sites are prominent features of sequence evolution in these data.

The phylogenetics of proteasome components have been studied by others who included entire gene families in their sampling [93,97]. Previous phylogenetic work on proteasome components analyzed amino acid sequences, which can be an effective means of determining phylogeny in highly divergent data. These analyses employ a variety of methods including MP, a non-model based algorithm, and NJ with a Poisson corrected distance. (The Poisson amino acid model is analogous to the JC69 nucleotide model and assumes that all changes between amino acids occur at the same rate and all amino acids are found in equal frequency.) The JTT amino acid model [98] is also used to calculate maximum likelihood scores of three preset fixed topologies. The JTT model is more suited to the analysis of divergent amino acid sequences and is based on substitution rates in a large sample of related proteins [98]. In these cases, the use of a particular model is reported but no tests were conducted to select from a suite of available

amino acid models [99,100]. Statistical theory is often utilized through model-based phylogenetics, but the fuller potential and benefits of statistical analysis remains unemployed by not considering recent advancements in model selection. Many other examples of using evolutionary models for phylogenetic reconstruction without statistically evaluating the fit of a model are widespread in the literature [72].

Another example of differing trees obtained with differing phylogenetic methodologies can be found in a study of antigen receptors by Richards and Nelson [101]. They used two methods, MP and NJ, to reconstruct the evolutionary history of members of the immunoglobulin superfamily of genes using amino acid sequences. For NJ distance calculations, they do not specify which model of evolution was used to estimate genetic distances or mention how that model was selected. However, in their analysis the model-based NJ method outperformed the MP because more monophyletic clades reflected current immune receptor classifications established by function [101]. Even with a model of evolution, strong bootstrap support for many of the nodes in their analysis is lacking. Such a lack of support or conflicting trees may be expected when the natural limitations of protein size and ancient divergence constrain the size and signal of the sequence alignment used for analysis. Also, similar structures and function in families of genes can cause convergence at the molecular level. Finally, the period following the gene duplications that create multigene families is often marked with increased substitution rates or varying levels of natural selection [102]. The temporal and often temporary change in evolutionary process makes phylogenetic analysis with stationary models of evolution more difficult.

Other data sets will have different properties that play an important role in determining the best-fit model, and population data collected from a single species presents unique obstacles for evolutionary analysis. The phylogenetic methods discussed here are designed for use on hierarchically ordered data (each sampling unit has only a single ancestor) such as the creation of two species from one. Their use on population-based sampling from a single species presents other difficulties which may further complicate analysis and create misleading results, even with correct use of statistically justified models [103]. For instance, in a population sample, sequences may not be related in a hierarchical manner (each unit has two ancestors (parents) in a sexually reproducing species). Further, processes at the population level, such as recombination, result in the problem that different parts of a DNA sequence have different evolutionary histories, and cannot accurately be represented by a single phylogenetic tree [104]. Use of different parts of recombining trees typically leads to different trees that may not be correlated, depending on the relationship of the sequences that exchanged genetic information [105]. Recombination also alters estimates of mutation rates, dating of evolutionary events, and estimates of among-site variation [106,107]. Extra effort should be taken when using phylogenetic methods to analyze data from a single species to avoid pitfalls introduced by population-level processes, and methods designed for this purpose should be employed [103].

5. Summary

The estimation of phylogenetic trees or genetic distances is a complex statistical problem in which elements such as rate of evolution, branch length, and tree topology are represented by parameters in a model [56]. A phylogenetic tree and model parameters should be considered a hypothesis of evolutionary relationships statistically supported by particular data. It is important to ensure that any conclusions from evolutionary genetic analysis be as strongly supported as possible by using statistically relevant models. Results obtained using arbitrarily selected models may easily be contradicted simply by using different models that lend support to different hypotheses [52,55]. When model-based methods are used, their performance is optimized when the best model is used [37], thereby lending more credibility to results obtained using statistically justified models. Some estimate of the fit of the model to the data should be calculated and used to select among available models rather than relying heavily on the robustness of the reconstruction method [72]. It is our position that statistical accuracy should not be sacrificed for the sake of ease or computational speed. Advancements in the statistics of model selection have already benefited every scientific

discipline that uses model-based analysis. Evolutionary genetic analysis is also experiencing similar progress from these advancements. New models and improved implementation, along with selection of models under a statistically rigorous framework will continue to enhance understanding of evolutionary patterns and processes underlying the variation found in genes of the immune system.

Acknowledgements

We would like to thank Louis Du Pasquier and Ashley Sparrow for helpful comments on an earlier version of this manuscript. Janae Bos and Matt Walters provided helpful editorial and digital imagery assistance. David Bos was supported by the Marsden Fund of New Zealand and a PhD scholarship from the University of Canterbury. David Posada is supported by the 'Ramón y Cajal' program of the Spanish government.

References

[1] Page RDM, Holmes EC. Molecular evolution: a phylogenetic approach. Cambridge: Blackwell Science; 1998.
[2] Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. Nature Rev Genet 2003;4:275–84.
[3] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 1987;4:406–25.
[4] Fitch WM. Toward defining the course of evolution: minimal change for a specific tree topology. Syst Zool 1970;20:406–16.
[5] Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 1981;17:368–76.
[6] Larget B, Simon D. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol Biol Evol 1999;16:750–9.
[7] Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 2003; 19:1572–4.
[8] Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 2001;294:2310–4.
[9] Brower A, DeSalle R, Vogler AP. Gene trees, species trees, and systematics: a cladisitic perspective. Ann Rev Ecol Syst 1996;27:423–50.
[10] Huelsenbeck JP, Crandall KA. Phylogeny estimation and hypothesis testing using maximum likelihood. Ann Rev Ecol Syst 1997;28:437–66.
[11] Nei M. Phylogenetic analysis in molecular evolutionary genetics. Ann Rev Genet 1996;30:371–403.
[12] Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst Biol 2001;50:525–39.
[13] Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. Sunderland, MA: Sinauer; 1996.
[14] Li W-H. Molecular evolution. Sunderland, MA: Sinauer; 1997.
[15] Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 1988;335:167–70.
[16] Yang Z. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 1996;11:367–72.
[17] Zhang J, Gu X. Correlation between the substitution rate and rate variation among sites in protein evolution. Genetics 1998;149:1615–25.
[18] Gu X, Zhang J. A simple method for estimating the parameter of substitution rate variation among sites. Mol Biol Evol 1997;14:1106–13.
[19] Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, editors. Evolving genes and proteins. New York: Academic Press; 1965. p. 97–166.
[20] Kimura M. Evolutionary rate at the molecular level. Nature 1968;217:624–6.
[21] Bromham L, Penny D. The modern molecular clock. Nature Rev Genet 2003;4:216–24.
[22] Ayala FJ. Molecular clock mirages. BioEssays 1999;21:71–5.
[23] Gillespie JH. The causes of molecular evolution. New York: Oxford University Press; 1991.
[24] Nei M, Gu X, Sitnikova T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci USA 1997;94:7799–806.
[25] Merritt TJS, Quattro JM. Evidence for a period of directional selection following gene duplication in a neutrally expressed locus of Triosephosphate Isomerase. Genetics 2001;159: 689–97.
[26] Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB. Estimating divergence times from molecular data on population genetic and phylogenetic time scales. Ann Rev Ecol Syst 2002;33:707–40.
[27] Sorhannus U, Van Bell C. Testing for equality of molecular evolutionary rates: a comparison between a relative-rate test and a likelihood ratio test. Mol Biol Evol 1999;16:849–55.
[28] Graur D, Martín W. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends Genet 2004;20:80–6.
[29] Huelsenbeck JP, Larget B, Swofford DL. A compound process for relaxing the molecular clock. Genetics 2000;154: 1879–92.
[30] Sanderson MJ. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol Biol Evol 1997;14:1218–32.
[31] Yoder AD, Yang Z. Estimation of primate speciation dates using local molecular clocks. Mol Biol Evol 2000;17: 1081–90.

[32] Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 1994;11:459–68.
[33] Huelsenbeck JP, Hillis DM. Success of phylogenetic methods in the four-taxon case. Syst Biol 1993;42:247–64.
[34] Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 1978;27: 401–10.
[35] Huelsenbeck JP. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of the maximum likelihood over neighbor joining. Mol Biol Evol 1995;12:843–9.
[36] Yang Z. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst Biol 1994;43:329–42.
[37] Huelsenbeck JP. Performance of phylogenetic methods in simulation. Syst Biol 1995;44:17–48.
[38] Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York, USA: Academic Press; 1969. p. 21–132.
[39] Kimura M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 1980;16:111–20.
[40] Kimura M. Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 1981;78:454–8.
[41] Tavare S. Some probabilistic and statistical problems in the analysis of DNA sequences. Lec Math Life Sci 1986;17: 57–86.
[42] Hasegawa M, Kishino H, Yano T. Dating the human–ape split by a molecular clock of mitochondrial DNA. J Mol Evol 1985;22:160–74.
[43] Felsenstein J. PHYLIP (Phylogenetic inference package). Seattle, WA: University of Washington; 1995.
[44] Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in Humans and Chimpanzees. Mol Biol Evol 1993;10:512–26.
[45] Sullivan J, Holsinger KA, Simon C. Among site rate variation and phylogenetic analysis of 12s rRNA in Sigmontine rodents. Mol Biol Evol 1995;12:988–1001.
[46] Yang Z. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 1993;10:1396–401.
[47] Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 1994;39:306–14.
[48] Pedersen A-MK, Wiuf C, Christiansen FB. A codon-based model designed to describe lentiviral evolution. Mol Biol Evol 1998;15:1069–81.
[49] Muse SV, Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 1994;11:715–24.
[50] Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 1994;11:725–36.
[51] Yang Z, Nielsen R, Goldman N, Pedersen A-MK. Codon substitution models for heterogeneous selection pressure and amino acid sites. Genetics 2000;155:431–49.
[52] Kelsey CR, Crandall KA, Voevodin AF. Different models, different trees: the geographic origin of PTLV-I. Mol Phylogent Evol 1999;13:336–47.
[53] Gu X, Li W-H. A general additive distance with time-reversibility and rate variation among nucleotide sites. Proc Natl Acad Sci USA 1996;93:4671–6.
[54] Buckley TR, Simon C, Chambers GK. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst Biol 2001;50:67–86.
[55] Cunningham CW, Zhu H, Hillis DM. Best-fit maximum likelihood models for phylogenetic inference: empirical tests with known phylogenies. Evolution 1998;52:978–87.
[56] Yang Z, Goldman N, Friday A. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst Biol 1995;44:384–99.
[57] Wakeley J. Substitution rate variation among sites and the estimation of transition bias. Mol Biol Evol 1994;11:436–42.
[58] Tajima F, Takezaki N. Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol Biol Evol 1994;11:278–86.
[59] Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition–transversion and G+C content biases. Mol Biol Evol 1992;9:678–87.
[60] Yang Z, Goldman N, Friday A. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol 1994;11:316–24.
[61] Sanderson MJ, Kim J. Parametric phylogenetics? Syst Biol 2000;49:817–29.
[62] Sullivan J, Swofford DL. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst Biol 2001;50:723–9.
[63] Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985;39:783–91.
[64] Hillis DM, Bull JJ. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst Biol 1993;42:182–92.
[65] Buckely TR, Cunningham CW. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. Mol Biol Evol 2002;19:394–405.
[66] Lemmon AR, Moriarity EC. The importance of proper model assumption in Bayesian Phylogenetics. Syst Biol 2004;53: 265–77.
[67] Wilcox TP, Zwickl DJ, Heath TA, Hillis DM. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Mol Phylogent Evol 2002;25:361–71.
[68] Suzuki Y, Glazko GV, Nei M. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc Natl Acad Sci USA 2002;99:16138–43.
[69] Simmons MP, Pickett KM, Miya M. How meaningful are Bayesian support values? Mol Biol Evol 2004;21:188–99.

[70] Takahashi K, Nei M. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol Biol Evol 2000;17:1251–8.
[71] Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer; 2002.
[72] Posada D, Crandall KA. Selecting the best-fit model of nucleotide substitution. Syst Biol 2001;50:580–601.
[73] Swofford DL. PAUP* phylogenetic analysis using parsimony (*and other methods). Version 4.0. Sunderland, MA: Sinauer; 1998.
[74] Kumar S, Tamura K, Jakobsen IB, Nei M. MEGA2: molecular evolutionary genetics analysis software. Tempe: Arizona State University; 2001.
[75] Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 2001;17:754–5.
[76] Rzhetsky A, Nei M. Tests of applicability of several substitution models for DNA sequence data. Mol Biol Evol 1995;12:131–51.
[77] Whelan S, Goldman N. Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol Biol Evol 1999;16:1292–9.
[78] Goldman N, Anderson JP, Rodrigo AG. Likelihood-based tests of topologies in phylogenetics. Syst Biol 2000;49: 652–70.
[79] Goldman N, Whelan S. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. Mol Biol Evol 2000;17:975–8.
[80] Posada D, Crandall KA. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci USA 2001;98:13757–62.
[81] Akaike H. A new look at the statistical model identification. IEEE Trans Autom Contr 1974;19:716–23.
[82] Suchard MA, Weiss RE, Sinsheimer JS. Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol 2001;18:1001–13.
[83] Huelsenbeck JP, Larget B, Alfaro ME. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol Biol Evol 2004;21:1123–33.
[84] Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6:461–4.
[85] Raftery AE. Hypothesis testing and model selection. In: Gilks WR, Richardson S, Spiegelhalter DJ, editors. Markov chain Monte Carlo in practice. London: Chapman & Hall; 1996. p. 163–87.
[86] Navidi WC, Churchill GA, von Haeshler A. Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol Biol Evol 1991;8:128–43.
[87] Goldman N. Statistical tests of models of DNA substitution. J Mol Evol 1993;36:182–98.
[88] Bollback JP. Bayesian model adequacy and choice in phylogenetics. Mol Biol Evol 2002;19:1171–80.
[89] Posada D, Crandall KA. MODELTEST: testing the model of DNA substitution. Bioinformatics 1998;14:817–8.
[90] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994; 22:4673–80.
[91] Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 1999;16:1114–6.
[92] Rzhetsky A, Sitnikova T. When is it safe to use an oversimplified substitution model in tree making? Mol Biol Evol 1996;13:1255–65.
[93] Takezaki N, Zaleska-Rutczynska Z, Figueroa F. Sequencing of amphioxus *PSMB5/8* gene and phylogenetic position of agnathan sequences. Gene 2002;282:179–87.
[94] Gu X, Li W-H. Estimation of evolutionary distances under stationary and nonstationary models of nucleotide substitution. Proc Natl Acad Sci USA 1998;95:5899–905.
[95] Huelsenbeck JP, Nielsen R. Variation in the pattern of nucleotide substitution across sites. J Mol Evol 1999;48: 86–93.
[96] Whelan S, Lio P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet 2001;17:262–72.
[97] Hughes AL. Evolution of the proteasome components. Immunogenetics 1997;46:82–92.
[98] Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. Comp Appl Bioscience 1992;8:275–82.
[99] Kishino H, Miyata T, Hasegawa M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 1990;31:151–60.
[100] Whelan S, Goldman N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 2001;18:691–9.
[101] Richards MH, Nelson JL. The evolution of vetebrate antigen receptors: a phylogenetic approach. Mol Biol Evol 2000;17: 146–55.
[102] Moore RC, Purugganan MD. The early stages of duplicate gene evolution. Proc Natl Acad Sci USA 2003;100:15682–7.
[103] Posada D, Crandall KA. Intraspecific gene genealogies: trees grafting into networks. Trends Ecol Evol 2001;16:37–45.
[104] Schierup MH, Hein J. Consequences of recombination on traditional phylogenetic analysis. Genetics 2000;156:879–91.
[105] Posada D, Crandall KA. The effect of recombination on the accuracy of phylogeny estimation. J Mol Evol 2002;54: 396–402.
[106] Satta Y, Kupferman H, Li Y-J, Takahata N. Molecular clock and recombination in primate MHC genes. Immunol Rev 1999;167:367–79.
[107] Schierup MH, Mikkelsen AM, Hein J. Recombination, balancing selection, and phylogenies in MHC and self-incompatibility genes. Genetics 2001;159:1833–44.
[108] Zharkikh A. Estimation of evolutionary distances between nucleotide sequences. J Mol Evol 1994;39:315–29.
[109] Rodriguez F, Oliver JF, Marin A, Medina JR. The general stochastic model of nucleotide substitution. J Theor Biol 1990;142:485–501.

J Mol Evol (2005) 60:221–228
DOI: 10.1007/s00239-004-0120-x

JOURNAL OF MOLECULAR EVOLUTION

Natural Selection During Functional Divergence to *LMP7* and Proteasome Subunit X (*PSMB5*) Following Gene Duplication

David H. Bos

School of Biological Sciences, University of Canterbury, Christchurch, New Zealand

Received: 16 April 2004 / Accepted: 9 September 2004 [*Reviewing Editor:* Dr. Rasmus Nielsen]

Abstract. The *LMP7* and *PSMB5* genes were created through an ancient gene duplication event of their ancestral locus. These proteins contain an active site of proteolysis, and *LMP7* replaces *PSMB5* as a component of the 20S proteasome after stimulation of cells by interferon-γ. Replacement of *PSMB5* by *LMP7* changes the profile of the products of 20S proteasome processing, predisposing digested peptides for transport to and display by the immune system. The purpose of this study is to investigate evolutionary forces influencing functional divergence between *LMP7* and *PSMB5* following duplication. Levels of synonymous and nonsynonymous substitution rates are estimated to infer differences in levels of natural selection. Estimates of substitution rates indicate that natural selection elevated rates of nonsynonymous substitution in *LMP7* following gene duplication, whereas *PSMB5* experienced an increase in substitution rate that was not likely due to diversifying natural selection following duplication. Following initial divergence, nearly neutral mutations have dominated gene evolution in both lineages. The *LMP7* gene locus provides a rare example of a protein with specialized function arising from duplication and divergence of a housekeeping protein by way of natural selection.

Key words: Subfunctionalization — MHC — Nonsynonymous substitutions — d_n — d_s — Molecular adaptation — AIC

Correspondence to: David H. Bos; *email:* dbos@purdue.edu

Introduction

Gene duplications are an important factor in molecular evolution and are a major mechanism through which proteins can assume new or specialized functions. Through duplications of genomes, chromosomes or genes, several copies of a gene may arise and later diverge to form multigene families (Li 1997; Ohno 1970). Duplication events can result in one of several outcomes, brought about by a combination of various competing mechanisms. For the vast majority of gene duplications, redundant gene loci will likely degenerate into nonfunctional pseudogenes due to the deleterious nature of most mutations and the initial low frequency of the haplotype (Lynch and Conery 2000; Walsh 1995). However, models of duplicate gene loss are difficult to reconcile with the relatively high numbers of duplicate loci found in the genomes of some model organisms (Hughes and Hughes 1993; Prince and Pickett 2002). Recently, theoretical work has focused on mechanisms that preserve duplicated genes from loss due to a null mutation.

The result of a gene duplication can be influenced by several evolutionary factors. When an ancestral gene has multiple roles, subfunctionalization theory predicts that both gene copies may be preserved and each assumes a different subset of ancestral functions (Force et al. 1999). Under this scenario, a gene copy may become unable to perform particular ancestral functions due to the deterioration of one or more regulatory regions, and the other gene copy is then preserved to carry out these functions. Gene duplicates can also be preserved due to purifying selection

that occurs because the protein has multiple domains or is part of a molecule with several subunits (Gibson and Spring 1998). In this case, point mutations may result in a stronger phenotype than a null mutation, and as long as gene duplicates are expressed, it is possible to be preserved as a subunit of a molecule due to purifying selection on the structure of the multiple domains or multisubunit molecule. Finally, it has been proposed that a gene copy can persist long enough to specialize or acquire a new function through the forces of positive (diversifying) natural selection (Hill and Hastie 1987). Divergence according to positive natural selection is often manifested through higher rates of amino acid substitution compared to the synonymous substitution rate in DNA. In fact, all the mechanisms of gene locus maintenance outlined above often involve a change in the rate of substitutions from the basal rate.

The period after a gene duplication is often marked by an interval of increased substitutions but the cause of this increased rate is debated. Initially, the rate increase was attributed to neutral mutations due to relaxed selective constraints on duplicate gene copies (Ohno 1970). More recently, a role for natural selection at the molecular level has been postulated (Hughes 1994). However, it is often problematical to distinguish the effects of neutral evolution and evolution by natural selection. Furthermore, natural selection may be difficult to demonstrate because it is likely to occur at only a few sites in a sequence and may exert influence for only a short amount of time and act differently on gene duplicates (Golding and Dean 1998; Yang 2001). Despite the difficulty characterizing older duplication events, they are of particular interest because of the high numbers of duplication events reported to occur in early vertebrate evolution (Gu et al. 2002; McLysaght et al. 2002; Ohno 1970). Some loci involved in duplications during early vertebrate evolution include genes of the adaptive immune system, many of which have related duplicates in paralogous regions of the genome (Abi Rached et al. 2002; Flajnik and Kasahara 2001).

The 20S proteasome is a vital housekeeping component of the cell and is responsible for the constant degradation of cellular proteins into short peptides and amino acids (Rock et al. 1994). The 20S proteasome is comprised of α and β subunits, of which some β subunits contain the active site of proteolysis (Arendt and Hochstrasser 1997; Heinemeyer et al. 1997; Seemuller et al. 1995). In most vertebrates, stimulation of cells by interferon-γ alters the biochemical profile of cleavage sites and size spectrum of peptide products (Boes et al. 1994; Driscoll et al. 1993), a result of the replacement of the three conventionally expressed active β subunits by closely related gene family members. The conventionally expressed subunits of the housekeeping proteasome, *X*, *Y*, and *Z* (human genome database coded as *PSMB5*, *PSMB6*, and *PSMB7*, respectively), are replaced by *LMP7* (*PSMB8*), *LMP2* (*PSMB9*), and *MECL1* (*PSMB10*), respectively (Coux 1996). When these three new subunits are in place, the action of the proteasome changes so that proteins are cleaved more frequently after hydrophobic residues and less after acidic residues (Gaczynska et al. 1994; Toes et al. 2001). These peptides are more effectively transported to the endoplasmic reticulum (ER) and loaded onto major histocompatibility complex (MHC) class I proteins that are displayed on the cell surface to cytotoxic T cells (Coux 1996; Rock et al. 2002).

Proteasome components *LMP7* and *LMP2* are found in the MHC regions of vertebrates that encodes the MHC class I genes and other functionally related genes of the immune system (Flajnik and Kasahara 2001). Proteasomes with these interferon-γ inducible subunits are called immunoproteasomes and have already been shown to be functionally distinct from the constitutively expressed proteasomes due to the incorporation of active β subunits *LMP2*, *LMP7*, and *MECL1* (Boes et al. 1994; Kesmir et al. 2003). The constitutive and interferon-γ inducible proteasome subunit pairs originate from duplication of the three ancestral loci (Hughes 1997). Linkage patterns and inferred homology have indicated that the three interferon-γ inducible forms were possibly created by simultaneous chromosomal duplication of the more ancient *PSMB5*, *–6*, and *–7* (Clark et al. 2000; Kasahara et al. 1996).

The mechanism of functional diversification since duplication is of particular interest, and two main classes of theories on the predominant form of diversification have been proposed. The main difference between paradigms of diversification is the emphasis placed on neutral mutation and drift versus natural selection (Wagner 2002; Zhang et al. 1998). Previous work on *PSMB5* and *LMP7* found evidence for an increased substitution rate in interferon-γ inducible proteasome subunits, possibly associated with acquisition of specialized (sub)function (Takezaki et al. 2002). Here, I build on the work of Takezaki et al. (2002) by further investigating the nature and cause of the elevated substitution rates in specific lineages.

Examples of positive selection operating to diversify the function of vertebrate gene families are common in the current literature but are mostly limited to loci involved in reproductive isolation or non-self-recognition. In this research, the focus is to study sequences of *PSMB5* and *LMP7* to elucidate molecular evolutionary forces causing nucleotide divergence. I specifically target the interval following the duplication of the proto-*PSMB5*/*LMP7* housekeeping gene and investigate substitution rates to infer the operation of diversifying natural selection.

Table 1. Candidate models used to estimate codon substitution rates

Model	fp[a]	rates1[b]	rates2[c]	Foreground	Parameter(s)
			One-rate model		
M0	50	1	1	None	ω_0
			Site-specific models		
M3 ($k = 3$)	54	1	3	None	$p_0\ p_1\ p_2$ $\omega_0\ \omega_1\ \omega_2$
M3 ($k = 2$)	52	1	2	None	$p_0\ p_1$ $\omega_0\ \omega_1$
			Branch-specific models (two -ratios)		
Br1	51	2	1	Branch 1	$\omega_0\ \omega_1$
Br2	51	2	1	Branch 2	$\omega_0\ \omega_1$
Br3	51	2	1	Branch 3	$\omega_0\ \omega_1$
Br4	51	3	1	Branch 4	$\omega_0\ \omega_1$
			Branch-site models (model B of Yang and Nielson [2002])		
MB-1	54	2	3	Branch 1	$p_0\ p_1\ p_2\ p_3$ $\omega_0\ \omega_1\ \omega_2$
MB-2	54	2	3	Branch 2	$p_0\ p_1\ p_2\ p_3$ $\omega_0\ \omega_1\ \omega_2$
MB-3	54	2	3	Branch 3	$p_0\ p_1\ p_2\ p_3$ $\omega_0\ \omega_1\ \omega_2$
MB-4	54	2	3	Branch 4	$p_0\ p_1\ p_2\ p_3$ $\omega_0\ \omega_1\ \omega_2$

[a]Number of free parameters including branch lengths.
[b]Number of rates among branches.
[c]Number of rates among codon sites.

Results indicate that this is a rare example of divergence of an essential housekeeping gene (the proto-*PSMB5*/*LMP7*) involving duplication and evolution under natural selection.

Materials and Methods

Data

Twenty-one sequences were downloaded from the GenBank database. These data are from a variety of vertebrate taxa and are comprised of *LMP7* and/or *PSMB5* coding regions. Sequences are primarily from vertebrate taxa because interest is in the evolution of the paralogs created after gene duplication. Sequence from a tunicate and a lancet were included in the data to isolate the period of evolutionary time prior to and following the putative duplication event and because evidence suggests that these organisms are close outgroups to *PSMB5* and *LMP7* (Takezaki et al. 2002). All sequences were aligned in Clustal X (Thompson et al. 1994) and DNA sequences were checked to ensure that the alignment preserved the coding frame. The resulting alignment consists of 585 nucleotides and is the same as that found by Takezaki et al. (2002). Testing for saturation was done using the index of substitution saturation (Xia et al. 2003). The index score (I_{SS}) is significantly lower than the critical score ($I_{SS.C}$) for these data ($I_{SS} = 0.49$, $I_{SS.C} = 0.72$; $P < 0.001$).

Substitution Rate Estimation

Models of codon evolution that estimate numbers of nonsynonymous (d_n) and synonymous (d_s) substitutions, and calculate $d_n/d_s = \omega$ rates are used to infer levels of natural selection (Nielsen and Yang 1998; Yang et al. 2000). Overall, eleven models are included as candidates to describe these data, and codon frequencies are estimated using the F3X4 option. Models are designated as being site-specific, branch-specific, or branch-site models. A simple one-rate model is described by Goldman and Yang (1994) and is included for direct comparison to other models and the possibility that the data cannot support inference from more complex models. Site-specific models are described by Yang et al. (2000) and branch-specific and branch-site models are laid out by Yang (1998) and Yang and Nielson (2002), respectively. For branch-specific and branch-site models, evolutionary lineages with the basal substitution rate are referred to as being in the "background" or having the background rate, and lineages targeted as having a different rate are said to be in the "foreground" of the tree topology (see Table 1 and Fig. 1).

The same branching topology reconstructed by Takezaki et al. (2002) is used here, and it indicates that duplication of ancestral proteasome components occurred after divergence of vertebrates from amphioxus and prior to the separation of gnathostomes and agnathans. Because substitution rates are not constant in this tree, I primarily use models that combine rate variation both in time and among codons to more accurately approximate these findings (Burnham and Anderson 2002). Variants of branch-specific or branch-site models focus on lineages in which natural selection may have operated for a short time. These branches include the ancestral lineage to all vertebrates, lineages immediately following the duplication event (including the ancestral lineage of *PSMB5* in jawed vertebrates), and the common ancestral lineage of *LMP7*. These lineages are a priori designated as foreground branches in various models because substitutions that occur shortly after a duplication or lineage divergence can have a large impact on the subsequent evolution of gene loci, and the relative forces directing these substitutions are of interest.

Model parameters were estimated using CODEML in the PAML package (Yang 2000), and empirical Bayes methods are used to identify which specific sites are likely to fall into various substitution rate categories, thus identifying sites with $\omega > 1$. Maximum likelihood estimation procedures are used because they have been shown to perform well for deeply divergent genes (Muse 1996). These methods have been used on older duplication events (Rodriguez-Trelles et al. 2003), and similar methods have been

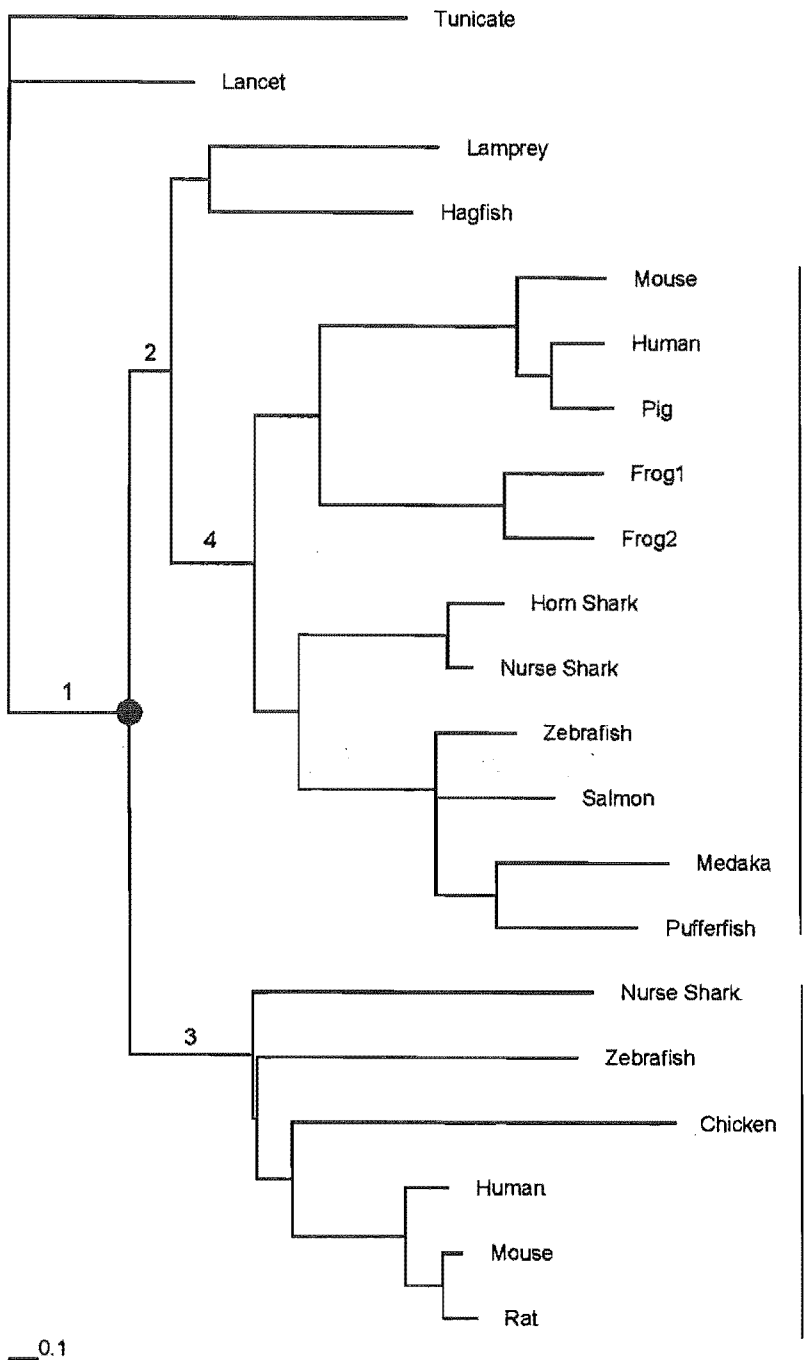


Fig. 1. Evolutionary relationships among taxa inferred from Takezaki et al. (2002) using *LMP7* and *PSMB5* sequences. Branch lengths shown are from the best-approximating model derived from current analyses. The putative duplication event is shown as a filled circle and numbered branches are those used as foreground branches. Accession numbers are AF449497 (lancet, *Branchiostoma lanceolatum*); X97729 (tunicate, *Botryllus schlosseri*); D64054 (hagfish, *Myxine glutinosa*); D64055 (lamprey, *Petromyzon merinus*); *PSMB5*: D29011 (human, *Homo sapiens*); AF060091 (mouse, *Mus musculus*); D45247 (rat, *Rattus rattus*); AB001935 (chicken, *Gallus gallus*); AF155578 (zebrafish, *Danio rerio*); D64058 (nurse shark, (*Ginglymostoma cirratum*); *LMP7*—BC001114 (human, *Homo sapiens*); AF059493 (pig, *Sus scrofa*); U22032 (mouse, *Mus musculus*); D44549, D44540 (African clawed frog, *Xenopus laevis*); D89725 (medaka, *Orizyias latipes*); AJ271723 (pufferfish, *Fugu rubripes*); AF184938 (salmon, *Salmo salar*); AF032390 (zebrafish, *Danio rerio*); D64057 (nurse shark, *Ginglymostoma cirratum*); AF363583 (horn shark, *Heterodontus franscisi*).

shown to be accurate at comparable divergence levels (Anisimova et al. 2001). Further, these data do not show signs of substitution saturation so that estimates of substitution rates should be reliable. It is also noted that estimating substitution rates of divergent genes may result in underestimation of substitutions, reducing the difference between d_n and d_s, making inference of selection more conservative (Suzuki and Nei 2001). Akaike's (1974) information criterion (AIC) is used to rank candidate models according to how well the model describes the patterns of substitutions in the data. Since AIC scores are relative measures, the model with the lowest score is taken as the benchmark, and the difference between the best score and all other candidate model scores (δAIC) is presented. Once models are ranked based on ability to describe the data, inference is drawn from parameter estimates of the best approximating model.

Results

Model Fitness and Selection

Maximum likelihood scores of the 11 candidate models varied depending on which factors were included. Generally, models that do not account for rate variation among sites fit the data more poorly than models that include parameters for variation among codons (Table 2). Likelihood scores of models that account for rate variation among both lineages and codons were much higher than those without among-site codon rate variation. Among branch-

Table 2. Results of ML optimization of substitution parameters and δAIC rank

Model	lnL	δAIC	Parameter values
MB-4	−6718.731	best	$p_0 = 0.242$, $p_1 = 0.586$, $p_2 = 0.050$, $p_3 = 0.122$ $\omega_0 = 0.186$, $\omega_1 = 0.013$, $\boldsymbol{\omega_2 = 999}$
M3 ($k = 3$)	−6720.760	4.06	$p_0 = 0.621$, $p_1 = 0.293$, $p_2 = 0.086$ $\omega_0 = 0.010$, $\omega_1 = 0.108$, $\omega_2 = 0.347$
MB-1	−6730.537	23.61	$p_0 = 0.014$, $p_1 = 0.035$, $p_2 = 0.278$, $p_3 = 0.673$ $\omega_0 = 0.184$, $\omega_1 = 0.015$, $\omega_2 = 0.00$
M3 ($k = 2$)	−6739.573	37.68	$p_0 = 0.700$, $p_1 = 0.300$ $\omega_0 = 0.014$, $\omega_1 = 0.183$
MB-3	−6738.075	38.69	$p_0 = 0.278$, $p_1 = 0.659$, $p_2 = 0.019$, $p_3 = 0.045$ $\omega_0 = 0.186$, $\omega_1 = 0.014$, $\boldsymbol{\omega_2 = 3.34}$
MB-2	−6739.482	41.50	$p_0 = 0.164$, $p_1 = 0.383$, $p_2 = 0.136$, $p_3 = 0.317$ $\omega_0 = 0.183$, $\omega_1 = 0.014$, $\omega_2 = 0.00$
Br4	−6940.957	438.45	$\omega_0 = 0.053$, $\boldsymbol{\omega_1 = 999}$
Br1	−6950.375	457.29	$\omega_0 = 0.054$, $\omega_1 = 0.000$
M0	−6953.428	461.39	$\omega_0 = 0.055$
Br3	−6952.639	461.82	$\omega_0 = 0.055$, $\omega_1 = 0.206$
Br2	−6953.215	462.97	$\omega_0 = 0.055$, $\omega_1 = 0.313$

specific and branch-site models, those having branch 4 in the foreground had the highest likelihood values. Overall, MB-4 had the highest likelihood, and besides M3($k = 3$), other models had much lower likelihood scores (Table 2). Likelihood values will always be better for more complex models, so AIC statistics are calculated from likelihood scores in order to more directly compare models relative to each other and estimate how well they approximate the data while taking into account model complexity.

Candidate models differed widely in δAIC scores, although a few models clustered closely together with similar scores (Table 2). Typically, models with δAIC scores < 2 are considered to have substantial support as a good approximation to the data and models with δAIC values > 10 have essentially no support from the data (Burnham and Anderson 2002). Among all candidate models, MB-4 was the best approximation to the data. Remaining branch-site models with other lineages in the foreground did not approximate relevant patterns of variation in the data as well and had larger δAIC scores. M3($k = 3$) had a low δAIC value and was the second-best approximation overall. Other branch-site and site-specific models have δAIC scores that are higher than 10 and have essentially no empirical support as a good approximation to the data. The δAICs of branch-specific models were more than an order of magnitude larger compared to models with variation among codons. The large differences seen in δAIC scores were also a feature of parameters estimated from different models.

Substitution Ratios

Parameter estimates varied widely with the model used. The best-approximating model allows for rate variation among sites and a rate shift in the common ancestral lineage to homologous copies of *LMP7* (branch 4). According to this model, the majority of sites (58%) are conserved in all branches of the tree, and 24% of sites are moderately conserved, indicating that rates vary across sites for these data. Approximately 17% of sites experienced a temporary shift that elevated substitutions above the basal rate in branch 4, and the d_n/d_s value is > 1 in this lineage (Tables 2 and 3). There are several substitutions on branch 4, and in the analysis by Takezaki et al. (2002), this lineage is strongly supported by a high bootstrap value. Despite the abundant numbers of total substitutions, too few synonymous changes are inferred on this branch to accurately estimate ω_2. Substitution rates that result from using other models also denote variation in substitution rates within lineages and among codon sites.

Models with lineages other than branch 4 in the foreground also indicated rate shifts. For all branch-site models, background branches were conserved, the majority of sites with $\omega < 0.02$. Both branch 3 and branch 4 had $\omega_2 > 1$ estimates in the foreground for sites that changes evolutionary rate. Other foreground branches had a decrease in d_n/d_s for sites that may have temporarily changed rates (Table 2). Site-specific models also estimated a variation in substitution rate among codons, with most being highly conserved and none with $\omega > 1$. Since these models average substitution rates across all lineages, a temporary elevation in nonsynonymous substitution rate would likely not result in $\omega > 1$. Nevertheless, variation in rates among codons is a relatively important aspect of evolution in these proteins, as site-specific models were a better fit to the data than branch-specific models. In addition, a direct comparison between each branch-site model and M3($k = 2$) is possible because these models are nested. δAIC score

Table 3. Codon sites with ω > 1 under model MB-4

$0.5 > p \geq 0.01$	$0.01 > p$
30	21
58	24
62	32
67	46
88	48
107	53
136	65
179	77
186	84
	87
	91
	99
	114
	125
	129

comparison reveals that adding a rate shift in branch 4 represents a substantial improvement in the fit of the model, whereas a rate shift in other branches tested here does not result in an improvement of the model.

Discussion

Model Fitness

The set of candidate models used here to investigate *PSMB5/LMP7* divergence since duplication allow the investigation of the relative roles of competing evolutionary forces through patterns of nucleotide substitution. The best-approximating model indicates that there is rate variation among codons and also a substantial rate shift in branch 4 of the tree topology. This shift results in an elevated nonsynonymous substation rate that is most easily explained by the operation of natural selection for a short period of time. Differences in population size or effectiveness of purifying selection may also lead to elevated nonsynonymous substitution rate, but overlapping taxonomic sampling between *LMP7* and *PSMB5* makes this explanation less likely. Results also indicate an increase in the rate of evolution in branch 3, but this model did not represent the data better than models with no temporary rate shift. Therefore, the evidence of natural selection in this branch is unconvincing because it does not improve the fit of the model. The ancestral lineage to *PSMB5* homologues probably did experience a rate shift, but it is not clear that it was due to natural selection. Subsequently, a model was used that included both branch 3 and branch 4 in the foreground, but this model did not have an improved fit to the data (*lnL*, −6738.726) compared to site-specific models, probably due to different sites with elevated rates of evolution in respective branches. Inference from parameters estimated from the best-approximating model is informative to identify prevalent forces acting to diversify duplicate gene loci.

Persistence of Duplicated Genes

Several theories explain the maintenance of duplicated loci that are not mutually exclusive, and the genomic organization and functions of *PSMB5* with *LMP7* indicate that multiple factors may have contributed. Immediately following the duplication event, both loci were likely expressed and incorporated into proteasomes, so it is possible that these loci were maintained in the gnathostome lineage due to negative selection as subunits of the proteasome (Gibson and Spring 1998). For duplicated proteasome components, a null mutation might not be devastating to the function of the proteasome because of redundancy, but a locus with a deleterious mutation to a site directly involved in proteolysis could reduce efficiency in a sizable fraction of the proteasomes in a cell. Therefore, purifying selection may have played an early role in preserving duplicate proteasome components in most descendant lineages.

Since *PSMB5* and *LMP7* currently have different expression patterns in tissue and timing (Akiyama et al. 1994), it is very likely that complementary degenerative mutations in regulatory regions also contributed to persistence of both loci, if not immediately, then shortly after duplication. When different regulatory regions are inactivated, then it is possible that the individual loci assumed distinct ancestral roles, although at this point it is likely that either locus could perform the suite of ancestral roles equally well. Nevertheless, the differences in expression patterns may have helped contribute to promoting differences between the loci and initiated processes that lead to functional diversification.

Functional Diversification

Results from the best-approximating model indicate that nonsynonymous substitution rates varied widely across sites in the sequences. The selective pressure to maintain structure and function differs in various parts of the *PSMB5/LMP7* proteins; however, there are several substitutions that are specific to just one subfamily (Hayashi et al. 1997; Kandil et al. 1996). Although agnathan sequences cluster with gnathostome *LMP7*, they lack *LMP7*-specific substitutions near the active site. Of particular interest is the cassette of codons in positions 27–31 that are near the active site and have radical substitutions in *LMP7* but not *PSMB5*. These sites may be identified evolving under natural selection for a time or as being "constant but different" sites that reflect differing roles of structure and function among the two subfamilies (Gribaldo et al. 2003).

In addition to the active site, the S1 pocket plays a crucial part in lysis by β subunits (Lowe et al. 1995). In *PSMB5/LMP7* this pocket plays a key role in specificity of cleavage due to steric interactions and biochemical properties of the pocket (Toes et al. 2001). In *PSMB5/LMP7*, the S1 pocket is comprised of amino acids from two adjacent subunits, and concerted movement and rotation of side chains in and near the S1 pocket in conjunction with substrate contact have been documented (Groll et al. 1997). Due in part to these complexities, the exact mechanisms responsible for functional divergence between constitutive and interferon-γ inducible forms of this β subunit remain unclear. Nevertheless, crystal structures have helped identify amino acids that comprise the reactive core (1, 17, 33), bind to the active site residues (129, 166, 168), form and determine the character of the S1 pocket (20, 31, 35, 45, 49, 53), and comprise additional residues in contact with substrate undergoing lysis in the proteasome (21, 47) (Groll et al. 1997; Unno et al. 2002).

Twenty-four sites are identified in these data as evolving for some time under natural selection (Table 3). Several of the selected residues in Table 3 are or are adjacent to residues identified above as involved in function and specificity of *LMP7*. Substitutions in *LMP7* near the above residues are positions 21, 30, 32, 46, 48, 125, and 129, and some of these sites also involve radical changes in amino acid properties. The positioning of these substitutions makes it possible that these sites have been involved with the functional divergence of these subfamilies by altering the stereochemistry of the S1 pocket. These changes may also alter the capacity for steric changes that occur in concert with substrate binding or the stoichiometry of the region surrounding the S1 pocket. Many substitutions taking place by natural selection occurred shortly after gene duplication, in branches ancestral to vertebrate *LMP7* homologues.

Branches 3 and 4 represent the ancestral lineage to *PSMB5* and *LMP7*, respectively, and have a higher rate of evolution than the basal rate found for the entire tree. The ω_2 estimate for branches 3 and 4 are both above 1. However, only an elevated rate of evolution in branch 4 improves the model fitness, an indication that natural selection was acting for a time to diversify ancient lineages of the *LMP7* locus, but the elevated substitution rate at the *PSMB5* locus was more likely due to relaxed selective constraints. The operation of natural selection in the ancestral branch of a functionally divergent gene locus supports the hypothesis that *LMP7* diverged in function through positively selected amino acid substitutions. Since that time, however, it is likely that nearly neutral evolution has dominated substitution rates in *LMP7*. The *LMP7* gene locus provides an example of a protein with specialized function arising from duplication, divergence, and natural selection of a housekeeping protein.

Acknowledgments. Bruce Waldman, Neil Gemmel, Martin Flajnik, Andrew DeWoody, and the DeWoody lab and peer referees provided valuable suggestions on the manuscript. Ziheng Yang is gratefully acknowledged for support with various versions of PAML. This work is supported by the Marsden Fund of New Zealand and by a University of Canterbury Doctoral Scholarship.

References

Abi Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H (2002) Evidence of *en bloc* duplication in vertebrate genomes. Nature Genet 31:100–105

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Contr 19:716–723

Akiyama K-y , Yokota K-y , Kagawa S, Shimbara N, Tamura T, Akioka H, Nothwang HG, Noda C, Tanaka K, Ichihara A (1994) cDNA cloning and interferon-γ down-regulation of proteasomal subunits X and Y. Science 265:1231–1234

Anisimova M, Beilawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18:1585–1592

Arendt CS, Hochstrasser M (1997) Identification of the yeast 20S proteasome catalytic centers and subunit interaction required for active-site formation. Proc Natl Acad Sci USA 94:7156–7161

Boes B, Hengel H, Ruppert T, Molthaup G, Koszinowski UH, Kloetzel P-M (1994) Interferon-γ stimulation modulates the proteolytic activity and cleavage site preference of 20S mouse proteasomes. J Exp Med 179:901–909

Burnham KP, Anderson DR (2002) Model Selection and Multimodel Inference: a practical information-theoretic approach Springer-Verlag, New York

Clark MS, Pontarotti P, Gilles A, Kelly A, Elgar G (2000) Identification and characterization of a β proteasome subunit cluster in the Japanese Pufferfish (*Fugu rubripes*). J Immunol 165:4446–4452

Coux O (1996) Structure and functions of the 20S and 26S proteasomes. Annu Rev Biochem 65:801–847

Driscoll J, Brown MG, Finley D, Monaco JJ (1993) MHC-linked *LMP* gene products specifically alter peptidase activities of the proteasome. Nature 365:262–264

Flajnik MF, Kasahara M (2001) Comparative genomics of the MHC: glimpses into the evolution of the adaptive immune system. Immunity 15:351–362

Force A, Lynch M, Pickett FB, Amores A, Yan Y-l , Postlethwait JH (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151:1531–1545

Gaczynska M, Rock KL, Spies T, Goldberg AL (1994) Peptidase activities of proteasomes are differentially regulated by the major histocompatibility complex-encoded genes for LMP2 and LMP7. Proc Natl Acad Sci USA 91:9213–9217

Gibson TJ, Spring J (1998) Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet 14:46–49

Golding GB, Dean AM (1998) The structural basis of molecular adaptation. Mol Biol Evol 15:355–369

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Gribaldo S, Casane D, Lopez P, Philippe H (2003) Functional divergence prediction from evolutionary analysis: A case study of vertebrate hemoglobin. Mol Biol Evol 20:1754–1759

Groll M, Ditzel L, Lowe J, Stock D, Bochtler M, Bartunik HD, Huber R (1997) Structure of the 20S proteasome from yeast at 2.4A resolution. Nature 386:463–471

Gu X, Wang Y, Gu J (2002) Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate gene evolution. Nature Genet 31:205–209

Hayashi M, Ishibashi T, Tanaka K, Kasahara M (1997) The mouse genes encoding the third pair of *B*-type proteasome subunits regulated reciprocally by INF-γ. J Immunol 159:2760–2770

Heinemeyer W, Fischer M, Krimmer T, Stachon U, Wolf DH (1997) The active sites of the eukaryotic 20S proteasome and their involvement in subunit precursor processing. J Biol Chem 272:25200–25209

Hill RE, Hastie ND (1987) Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature 326:96–99

Hughes AL (1994) The evolution of functionally novel proteins after gene duplication. Proc R Soc Lond B 256:119–124

Hughes AL (1997) Evolution of the proteasome components. Immunogenetics 46:82–92

Hughes MK, Hughes AL (1993) Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. Mol Biol Evol 10:1360–1369

Kandil E, Namikawa C, Nonaka M, Greenberg AS, Flajnik MF, Ishibashi T, Kasahara M (1996) Isolation of low molecular mass polypeptide complimentary DNA clones from primitive vertebrates. J Immunol 156:4225–4253

Kasahara M, Hayashi M, Tanaka K, Inoko H, Sugaya K, Ikemura T, Ishibashi T (1996) Chromosomal localization of the proteasome Z subunit gene reveals an ancient chromosomal duplication involving the major histocompatibility complex. Proc Natl Acad Sci USA 93:9096–9101

Kesmir C, van Noort V, de Boer RJ, Hogeweg P (2003) Bioinformatic analysis of functional differences between the immunoproteasome and the constitutive proteasome. Immunogenetics 55:437–449

Li W-H (1997) Molecular evolution. Sinauer, Sunderland, MA

Lowe J, Stock D, Jap B, Zwickl P, Baumeister W, Huber R (1995) Crystal structure of the 20S proteasome from the archeon *T. acidophilum* at the 3.4. Å resolution. Science 268:355–539

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

McLysaght A, Hokamp K, Wolfe KH (2002) Extensive genomic duplication during early chordate evolution. Nature Genet 31: 200–204

Muse SV (1996) Estimating synonymous and nonsynonymous substitution rates. Mol Biol Evol 13:105–114

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and application to the HIV-1 envelope gene. Genetics 148:929–936

Ohno S (1970) Evolution by gene duplication. Springer-Verlag, Berlin

Prince VE, Pickett FB (2002) Splitting pairs: the diverging fates of duplicated genes. Nature Rev Genet 3:827–837

Rock KL, Gramm C, Rothstein L, Clark K, Stein R, Dick L, Hwang D, Goldberg AL (1994) Inhibitors of the proteasome block the degradation of most cell proteins and the generation of peptides presented on MHC class I molecules. Cell 78:761–771

Rock KL, York IA, Saric T, Goldberg AL (2002) Protein degradation and the generaton of MHC class I-presented molecules. In: Dixon FJ (ed) Advances in immunology. Academic Press, San Diego, CA, pp 1–70

Rodriguez-Trelles F, Tarrio R, Ayala FJ (2003) Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the xanthine dehydrogenase gene. Proc Natl Acad Sci USA 100:13413–13417

Seemuller E, Lupas A, Stock D, Lowe J, Huber R, Baumeister W (1995) Proteasome from *Thermoplasm acidophilum:* A threonine protease. Science 268:579–582

Suzuki Y, Nei M (2001) Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Mol Biol Evol 18:2179–2185

Takezaki N, Zaleska-Rutczynska Z, Figueroa F (2002) Sequencing of amphioxus *PSMB5/8* gene and phylogenetic position of agnathan sequences. Gene 282:179–187

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Toes REM, Nussbaum AK, Degermann S, Schirle M, Emmerich NPN, Kraft M, Laplace C, Zwinderman A, Dick TP, Muller J, Schonfisch B, Schmid C, Fehling H-J, Stevanovic S, Rammensee H-G, Shild H (2001) Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. J Exp Med 194:1–12

Unno M, Mizushima T, Morimoto Y, Tomisugi Y, Tanaka K, Yasuoka N, Tsukihara T (2002) The structure of the mammalian 20S proteasome at 2.75 A resolution. Structure 10:609–618

Wagner A (2002) Selection and gene duplication: a view from the genome. Genome Biol 3:1012.1–1012.3

Walsh JB (1995) How often do duplicated genes evolve new functions? Genetics 139:421–428

Xia X, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. Mol Phylogent Evol 26:1–7

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang Z (2000) Phylogenetic analysis by maximum likelihood (PAML). University College London, London

Yang Z (2001) Adaptive Molecular evolution. In: Balding DJ, Cannings C, Bishop M (eds) Handbook of statistical genetics. John Wiley and Sons, New York, pp 327–350

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908–917

Yang Z, Nielsen R, Goldman N, Pedersen A-MK (2000) Codon substitution models for heterogeneous selection pressure and amino acid sites. Genetics 155:431–449

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA 95:3708–3713

Evolution by Recombination and Transspecies Polymorphism in the MHC Class I Gene of *Xenopus laevis*

David H. Bos[1] *and Bruce Waldman*

School of Biological Sciences, University of Canterbury, Christchurch, New Zealand

The patterns of major histocompatibility complex (MHC) evolution involve duplications, deletions, and independent divergence of loci during episodes punctuated by natural selection. Major differences in MHC evolution among taxa have previously been attributed to variation in linkage patterns of class I and class II MHC genes. Here we characterize patterns of evolution in the MHC class Ia gene of *Xenopus laevis* in terms of polymorphism, recombination, and extent of transspecies polymorphism. We also compare these patterns to see if a correlation exists with linkage or separation of the MHC class I and class II regions as seen in amphibians and teleost fishes. In *X. laevis*, we find high levels of polymorphism. Also, genetic exchange is relatively frequent and occurs in intron II, reshuffling allelic forms of exons 2 and 3. Evolutionary relationships among class I alleles show an intermingling of alleles from divergent *Xenopus* species rather than a species-specific clustering. Results indicate that the patterns of evolution are similar to those found in salmonid fishes and are different from the mode of evolution seen in primates. Similar patterns of class Ia evolution in salmonid fishes and *X. laevis* suggest that nonlinkage of class I and class II regions alone is insufficient to explain some patterns of MHC evolution in salmonids.

Introduction

Major histocompatibility complex (MHC) class I and class II genes are found in all gnathostomes and encode structurally similar proteins that present antigenic peptides to T lymphocytes. Class II proteins are expressed mainly on specialized antigen-presenting cells and primarily function to bind peptides derived from extracellular pathogens. Class I proteins are expressed in almost all cells and are involved in monitoring the internal environment of the cell for foreign, mutated, or misfolded proteins. Class I genes comprise classical (class Ia) and nonclassical (class Ib) loci which differ in polymorphism, structure, function, and expression pattern. Aside from their important role in the immune system, the MHC genes are of particular interest because of their patterns of genetic diversity.

Class I and class II alleles have several exceptional features. They exhibit very high levels of allelic diversity and amino acid polymorphism (Parham and Ohta 1996). Some MHC allelic lineages also exhibit unusual longevity which predates the formation of species (Figueroa, Gunther, and Klein 1988; Lawlor et al. 1988). As a result, "transspecies polymorphisms" exist, whereby some MHC alleles from separate species cluster together in phylogenetic analysis to the exclusion of alleles from within the same species. Variations in linkage patterns, order of gene loci, and the number of gene family members resulting from tandem duplications have also been observed (Kelley, Walter, and Trowsdale 2005). Many of the features of the MHC have been attributed to the forces of balancing selection acting at the molecular level (Hughes and Yeager 1998).

In the typical pattern of MHC evolution, class I and class II families of genes evolve differently from each other, each with its own rate of duplication, divergence, longevity, and pattern of recombination. Specifically, class II transspecies polymorphisms extend further back in evolutionary history than in class I lineages (Bontrop et al. 1999; Vogel et al. 1999). Class II loci also have higher levels of polymorphism than class I loci. Additionally, allelic recombination has been characterized by the intralocus exchange of small minicassettes of nucleotides that can occur throughout the length of the gene and has no single prominent breakpoint (A. L. Hughes, M. K. Hughes, and Watkins 1993; Jakobsen, Wilson, and Easteal 1998). Until more recently, it was not known whether these well-established patterns of MHC evolution were also found in species other than the mammalian model organisms.

Despite many elements of conserved structure and function, MHC evolution in nonmammalian taxa differs from well-defined norms. For instance, class II gene evolution in birds differs from that in mammals; the loci have a relatively recent origin in birds (Edwards, Wakeland, and Potts 1995; Hess and Edwards 2002). In bony fishes, the class I and class II genes, linked in a common region in all other vertebrates, are found on separate chromosomes (Bingulac-Popovic et al. 1997). However, this appears to be a derived trait in bony fishes because both class I and class II genes of sharks are closely linked (Ohta et al. 2000). Class Ia lineages in salmonid fishes share transspecies polymorphism among divergent taxa, whereas class II alleles cluster in a species-specific manner, just opposite the pattern in mammalian model organisms (Shum et al. 2001). Salmonid fishes also have much higher levels of class Ia than class II polymorphism. In general, recombination plays a more prominent role in teleost class Ia evolution than in mammalian benchmarks (Shum et al. 2001; Consuegra et al. 2005). Intragenic recombination in salmonids typically involves entire exons, and a prominent breakpoint for genetic exchange is easily identifiable. Authors have speculated that these patterns of evolution might be due to the nonlinkage of class I and class II loci as seen in bony fishes (Shum et al. 2001).

Xenopus laevis, the African clawed frog, is the first ectothermic vertebrate from which class I proteins were isolated (Flajnik et al. 1984). In this species, there is a single

[1] Present address: Department of Forestry and Natural Resources, Purdue University.

Key words: transspecies evolution, recombination, exon shuffling, Akaike information criterion (AIC), *Xenopus*, MHC.

E-mail: dbos@purdue.edu.

Mol. Biol. Evol. 22(12):1–7. 2005
doi:10.1093/molbev/msj016
Advance Access publication Month XX, XXXX

MHC class Ia locus with diploid inheritance patterns (Shum et al. 1993). This locus is highly polymorphic compared to mammals but is similar to the variation found in salmonid fishes. Another unusual aspect of *Xenopus* class Ia is the existence of two ancient allelic lineages (Flajnik et al. 1999). These lineages are very distinct as alleles belonging to different lineages are as divergent as MHC alleles from mouse and human. Linkage of class Ia and class II genes in *X. laevis* indicates a single MHC genomic region like many vertebrates, but unlike bony fishes (Nonaka et al. 1997).

Our motivation is to ascertain patterns of MHC class I evolution, by characterizing recombination, polymorphism, and extent of transspecies evolution in the class Ia gene of wild-caught *Xenopus* frogs. Previously established differences in the linkage of class I and class II regions between *Xenopus* and salmonid fishes also allow us to compare results and interpret these in light of hypotheses regarding the influence of linkage on MHC evolution. Differing patterns of class Ia evolution among these taxa would support the hypothesis that the mode of MHC evolution seen in salmonid fishes might be due to the nonlinkage of class I and class II genes. This mode of evolution was reported by Shum et al. (2001) and is distinguished by ancient class I lineages, high levels of polymorphism, and frequent recombination between the peptide-binding region (PBR)–coding exons. However, class Ia evolution in *X. laevis* that is similar to that of fishes supports the notion that nonlinkage of class Ia and class II genes alone may be insufficient to explain the mode of class Ia evolution reported for salmonids.

Materials and Methods

Data Collection

We extracted total RNA from *X. laevis* blood samples using the TRIzol protocol following the manufacturer's recommendations (Invitrogen). Total RNA was used to make cDNA in a Superscript One-Step reverse transcriptase–polymerase chain reaction (PCR) kit (Invitrogen), and first-strand synthesis was performed at 55°C for 25 min. Immediately after the first-strand synthesis, PCR was employed on cDNA with primers designed to amplify exons 2–4 of the MHC class Ia gene (forward primer: 5′-GTCACTCCCT-GCGYTAYTAT-3′ and reverse primer: 5′-TTTCTCCTT-CAGGCTGCTGT-3′). Primers were designed using Primer3 (Rozen and Skaletsky 2000) from known *X. laevis* sequences (Flajnik et al. 1999). The thermal profile used to amplify MHC fragments was optimized to minimize the occurrence of in vitro recombination (Judo, Wedel, and Wilson 1998). We cloned PCR products into the pCR 4 TOPO TA plasmid following the manufacturer's recommendations (Invitrogen), and recombinant DNA was transformed into TOP-10 *Escherichia coli* cells. *Escherichia coli* cells were plated onto LB agar and grown overnight at 37°C after which 6–10 individual colonies were picked and grown in LB broth at 37°C for 16 h.

A total of 5 ml of LB broth per cell matrix was removed, and plasmid DNA was extracted using alkaline lysis minipreps (Sambrook, Fritsch, and Maniatis 1989). We sequenced the MHC insert in both directions using BigDye v3.1 chemistry and an ABI 3730 automated sequencer. ABI trace files were edited using Bioedit (Hall 1999), and sequences were aligned using ClustalW (Thompson, Higgins, and Gibson 1994). Eleven new sequences were isolated from 11 *X. laevis* chromosomes; these sequences were independently verified from two to six separate colonies (GenBank accession numbers: DQ149596–DQ149606). Additional sequences were obtained but were not recovered multiple times and were excluded from the following analyses. New sequences were added to other known class Ia sequences of exons 2, 3, and 4 from frogs (GenBank accession numbers—*Xenopus tropicalis*: AY204558, AY204559 [*X. tropicalis* is also termed *Silurna tropicalis* but is listed herein as a congeneric *Xenopus* species due to strong monophyly of this species with other *Xenopus* {Evans et al. 2004}]; *Xenopus ruwenzoriensis*: AF497525–AF497528; and *X. laevis*, *Rana pipiens*, and a laboratory-bred interspecies hybrid of *X. laevis-Xenopus gilli*: AF185579–AF185588).

Statistical Analysis

We used various statistical methods to investigate evolutionary relationships and intragenic recombination. The program Maxchi was used to detect the occurrence of recombination events within *X. laevis* samples because it performs well in simulations and had a low false error rate (Posada and Crandall 2001*c*). The program RDP2 (Martin, Williamson, and Posada 2005) characterized intragenic recombination by identifying breakpoints and alleles created by recombination. The P value of significant differences used to infer recombination was set at 0.000005 with a window size of 20 nt to minimize the false-positive error rate (Martin, Williamson, and Posada 2005). To estimate population parameters (i.e., mutation [$\theta = 4N\mu$] and recombination [$\rho = 4Nr$]), we used the maximum likelihood (ML) method implemented in LDhat because this method uses a finite-sites model appropriate for highly divergent sequences (McVean, Awadalla, and Fearnhead 2002).

Evolutionary relationships were reconstructed among known *Xenopus* MHC class Ia sequences. Prior to phylogenetic analysis, we tested for the loss of information in these data due to saturation using the index of substitution saturation (I_{SS}) (Xia et al. 2003). We compared these values with the symmetric critical index of substitution saturation ($I_{SS.C}$) scores. For this procedure, we estimated the proportion of invariable sites and tested each data partition (see below) separately; each test included all sequences ($n = 27$). We used model-based algorithms to investigate evolutionary relationships among sequences (Bos and Posada 2005). The best approximating model of nucleotide evolution for these data was determined using Akaike's information criterion (AIC) (Akaike 1974). ML scores of candidate models were calculated using PAUP* 4.0 (Swofford 1998) and AIC scores computed in Modeltest (Posada and Crandall 1998). Employing the best approximating model, genetic distance and phylogenetic relationships were estimated using ML optimization.

A traditional bifurcating phylogenetic tree may not accurately represent evolutionary relationships among a population sample because of genetic exchange (Posada and Crandall 2001*a*). Therefore, we reconstructed separate trees by partitioning these data into congruent segments with

Table 1
***Xenopus laevis* Recombinant Sequences at the *UAA* Locus**

Recombinant Sequence	Nucleotide Breakpoint	Potential Parent Sequence
Xela r	230	*Xela *08*/unknown
*Xela *03*	328	*Xela *06/Xela *11*
Xela j	252	*Xela *05/Xela *06*
*Xela *06*	252	*Xela *07/Xela *02*
*Xela *08*	252	*Xela f/Xela *02*
*Xela *11*	252	*lg-a/c1/Xela *02*

Table 2
Population Parameters of *Xenopus laevis* Nucleotide Sequences (n = 19 samples chromosomes)

Protein Domain	α1	α2	α3	Total
Number of nucleotide sites	255	279	247	781
Number of variable sites	105	75	29	209
Number of haplotypes	16	18	16	19
Per site nucleotide diversity (π)	0.165	0.091	0.022	0.089
Per locus mutation rate (θ)	n.d.	n.d.	n.d.	46.08
Per locus recombination rate (ρ)	n.d.	n.d.	n.d.	66.39

NOTE.—n.d., not determined.

shared evolutionary history on either side of a putative recombination breakpoint. ML (Felsenstein 1981) and Neighbor-Joining (NJ; Saitou and Nei 1987) reconstructions were performed to test hypotheses regarding lineage assortment among species. ML trees were compared to a range of a priori topologies corresponding to different levels of trans-species lineage sharing among taxa. Tree comparisons were performed using Shimodaira-Hasegawa tests (Shimodaira and Hasegawa 1999; Shimodaira 2002) implemented in the program package Consel (Shimodaira and Hasegawa 2001). One thousand bootstrap pseudoreplications were used to estimate support for nodes in the NJ tree, and >50% bootstrap support in resulting topologies is shown.

Results

Data

Overall, the data (including *Rana* outgroup samples) consist of 27 sequences and have 397 polymorphic sites out of a total of 781 nt; on average, alleles differ by 99.30 nt. The $I_{SS.C}$ value represents the I_{SS} value beyond which data fail to recover a true phylogenetic tree. The α1 partition is 255 bp, and the α2/α3 domain partition is 526 bp; the respective I_{SS} values are 0.261 and 0.228. The $I_{SS.C}$ values are 0.682 and 0.715, respectively, and these are significantly larger than the respective I_{SS} scores ($P < 0.001$).

Population Parameters

Intragenic recombination plays a prominent role in *X. laevis* MHC evolution and is responsible for the creation of a number of new alleles. Among *X. laevis* sequences (including the *X. laevis-X. gilli* hybrids), the number of alleles created through recombination is at least 6 out of 19, over 30% of alleles in this data set (table 1). The parameters of RDP2 were set conservatively to avoid false-positive identification of recombination events, so this number represents a minimum number of recombinant alleles. Estimation of the population parameters shows relatively high mutation and recombination rates (table 2). Estimated values indicate that past mutation and recombination events both operate on the same scale and play a major role in shaping variation at this locus.

The recombination breakpoint also is shared among alleles, indicating that recombination is not free. The breakpoint of the recombinant alleles indicates the likely cross-over location of the genetic exchange, which commonly occurs in intron II in these data. The size of the DNA fragment involved in the exchanges typically encompasses about the first 255 nt or the entire α1 domain–coding exon (tables 1 and 2). This type of recombination leads to intra-locus allelic "exon shuffling" that creates new arrangements of existing variation in the PBR. This pattern is very different from that seen in humans, where genetic exchange involves much smaller fragments.

Some other trends in the pattern of recombination in *X. laevis* are noteworthy. For instance, some alleles are involved in genetic exchange more often than others. The occurrence of genetic exchange often involves *Xela-UAA*02* in these data. This allele is involved in creating at least three new alleles; *Xela-UAA*06* could be a parent sequence for two additional recombinant alleles. A bias in alleles involved in recombination has also been detected in salmonid fishes (Shum et al. 2001). Inspection of recombinants reveals that two recombinant alleles (alleles *UAA*06* and *UAA*11*) have identical α1 domain sequences, but the α2 and α3 domains of these two alleles are different. Finally, the formation of recombinant alleles is not restricted to closely related alleles as two highly divergent alleles can recombine (e.g., UAA^{J} and *UAA*02*). The apparent ongoing genetic exchange results in a high level of recombination that is likely to affect the evolutionary relationships among alleles and different domains of alleles.

Evolutionary Relationships

The evolutionary relationships among MHC class Ia alleles in *Xenopus* species were determined using ML estimation of genetic distances and NJ topology reconstruction. Tree reconstruction was done separately on two segments of the sequence, partitioning the nucleotide sequence fragment coding the α1 domain as one segment and the α2 and α3 domains together as the other segment. This partition was chosen to maximize the detection of different evolutionary histories due to genetic exchange and provides a means for confirming the presence of recombination in this data set. In the phylogenetic trees of the α1 and α2/α3 domains, recombinant alleles identified with the program RDP were found to be in different clades. These alleles moved across nodes with >50% bootstrap support to associate with different sets of alleles in each tree reconstruction. The translocation of alleles to different parts of the tree topology is consistent with patterns of recombination detected with RDP2.

The best approximating model selected for α1 domains differs from that chosen for the α2/α3 sequences (table 3). The best-fitting model for the α1 domain sequence evolution is TIM + Γ (Posada and Crandall 2001*b*); for the

Table 3
Best Approximating Model Parameters for Data Partitions

MHC Domain	α1	α2/α3
Selected model	TIM + Γ	K81uf + Γ
Base frequency		
A	0.328	0.276
T	0.201	0.215
G	0.253	0.295
C	0.218	0.214
Substitution rates		
A↔C	1.00	1.00
A↔G	2.43	1.89
A↔T	1.41	0.75
C↔G	1.41	0.75
C↔T	3.89	1.89
G↔T	1.00	1.00
Rate variation		
α parameter	0.576	0.466

α2/α3 sequence fragment, the K81uf + Γ (Kimura 1981) model was the best fit. Differences between data partitions are also found in tree topology, although the conservative Shimodiara-Hasagawa test indicates that the trees are not significantly different ($P > 0.05$).

The topology showing relationships among α1 domain sequences shows mixing of alleles from different species (fig. 1). Pairs of alleles from a species form well-supported groups in some cases, but both *X. tropicalis* and *X. ruwenzoriensis* α1 domains are intermingled together with *X. laevis* sequences. One group of *X. laevis* alleles is closely related and forms a tight cluster that has 100% bootstrap support; this group comprises five recombinant sequences. Most well-supported clades are near the tips of the tree and are comprised of only a few sequences; branches in the more basal parts of the tree are typically shorter than many terminal branches.

The evolutionary relationships reconstructed for α2/α3 domain sequences were different in some ways from the α1 domain sequence topology (fig. 2). For instance, some alleles segregated by species rather than being intermingled. In this tree, *X. tropicalis* alleles are basal in the topology and paraphyletic with respect to a clade of *X. laevis*, *X. gilli*, and *X. ruwenzoriensis* alleles. This topology establishes an incomplete separation of *X. tropicalis* alleles and monophyly of other sequences of the ingroup. All *X. ruwenzoriensis* alleles form a monophyletic cluster nested within a larger clade of *X. laevis* and *X. gilli*. There is a closely related group that forms a tight cluster similar to that seen in the α1 domain tree, but the cluster is comprised of different sequences and contains only one recombinant allele. This tree also has a mixture of both long and short terminal branches, but compared to the α1 domain tree, branches on the α2/α3 tree are much shorter.

Several a priori hypotheses were compared to the ML and NJ topologies for both the α1 and α2/α3 sequence partitions. A priori hypotheses are not exhaustive but are designed to gauge the level of transspecies allelic sharing among progressively more distantly related species. For instance, transspecies polymorphism could be confined to two relatively closely related species, such as *X. laevis* and *X. ruwenzoriensis*, or it may extend to more distantly related taxa, such as *X. laevis* and *X. tropicalis*. Hypotheses are designed to gauge similarity of the observed level of

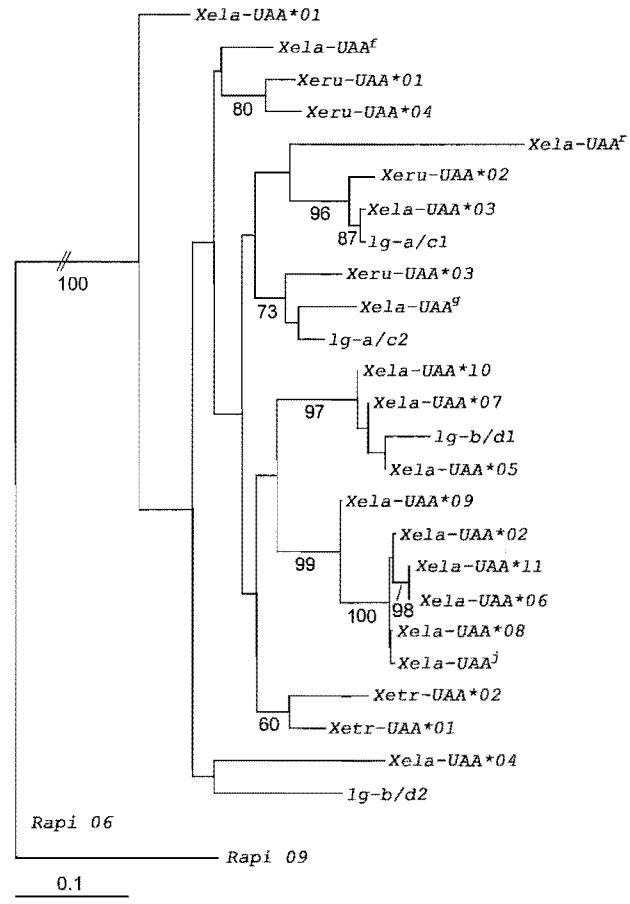


FIG. 1.—Evolutionary relationships of *Xenopus* class Ia sequences using α1 domain sequences. Numbers indicate bootstrap support for nodes. All branches shown to the scale at the bottom left of the figure, except the branch leading to the outgroup, which was shortened for graphical clarity of the remaining branches of the tree. *Rapi*, *Rana pipiens*; *Xela*, *Xenopus laevis*; *Xetr*, *Xenopus tropicalis*; *Xeru*, *Xenopus ruwenzoriensis*; and *lg*, *X. laevis-Xenopus gilli* laboratory hybrid.

transspecies polymorphism with constraints that represent various combinations of species that share alleles. Representative hypotheses include (1) no transspecies polymorphisms or reciprocal monophyly of species ((*X. laevis*),(*X. ruwenzoriensis*),(*X. tropicalis*),(*R. pipiens*)), (2) unconstrained *X. laevis* (*X. laevis*, (*X. ruwenzoriensis*),(*X. tropicalis*),(*R. pipiens*)), (3) transspecies polymorphism among two closely related species or unconstrained *X. laevis* and *X. ruwenzoriensis* (*X. laevis*, *X. ruwenzoriensis*, (*X. tropicalis*),(*R. pipiens*)), and (4) monophyletic *X. laevis* ((*X. laevis*), *X. ruwenzoriensis*, *X. tropicalis*, (*R. pipiens*)). Note that the ML tree serves to represent another hypothesis, namely, that transspecies polymorphism can occur throughout the genus *Xenopus*. Results of the Shimodaira-Hasegawa tests for both data partitions show that there is no significant difference between the ML and NJ trees (table 4). However, hypotheses 1 and 4 are significantly different from unconstrained optimal trees for both data partitions. For the α1 domain, hypothesis 2 is also significantly different from the optimal tree. For the α2/α3 data partition, tree constraints for hypotheses 2 and 3 resulted in the same phylogenetic reconstruction. Only hypothesis 3, where the only

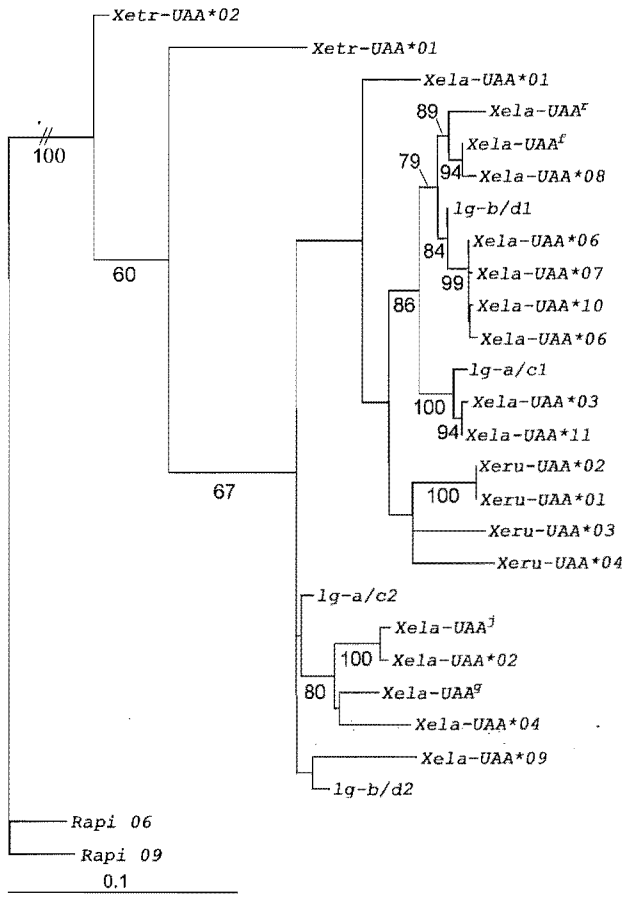


FIG. 2.—Evolutionary relationships of *Xenopus* class Ia sequences with *Rana* outgroup using α2 and α3 domain sequences. Numbers indicate bootstrap support for nodes. All branches scaled as in figure 1.

constraint is placing *R. pipiens* as an outgroup to *Xenopus*, is not significantly different from the unconstrained ML and NJ topologies for all data partitions.

Discussion

Recombination

Intragenic recombination plays a prominent role in creating and maintaining diversity in the class Ia gene of *X. laevis*. Here, a clear pattern of reticulate evolution is characterized by high rates and phylogenetic inconsistencies between the α1 domain and sites in the remaining 3′ exons of the gene. The nature of *X. laevis* recombination events is in contrast to the reticulate evolution seen among alleles of humans for three class I loci. Also, the pattern seen here would not be expected from in vitro chimera formation, where random breakpoints would be expected; therefore, we interpret this result to be due to in vivo allelic recombination. Although we have identified one recombination breakpoint in these sequences, the location of the other endpoint remains uncertain. However, given strong linkage disequilibrium with functionally related immunoproteasome (LMP) and transporter (TAP) loci (Namikawa et al. 1995; Ohta et al. 2003), genetic exchange events are expected to be much smaller than the distance separating these loci (Andolfatto and Nordborg 1998).

Table 4
Statistical Comparison of Phylogenetic Hypotheses for *Xenopus* Class Ia Alleles

Topology	ln L	δ ln L	S-H[a]	W S-H[b]
α1 Domain trees				
ML	−1844.420	—	—	—
3	−1844.905	0.5	0.850	0.812
NJ	−1858.748	14.3	0.491	0.281
2	−1920.962	76.5	0.000	0.000
4	−2009.004	164.6	0.000	0.000
1	−2058.469	214.0	0.000	0.000
α2/α3 Domain trees				
ML	−3003.147	—	—	—
3	−3003.382	0.2	0.848	0.861
2	−3003.382	0.2	0.848	0.877
NJ	−3003.710	0.6	0.743	0.727
1	−3148.052	144.9	0.000	0.000
4	−3168.448	165.3	0.000	0.000

[a] Shimodaira-Hasegawa (1999) test.
[b] Weighted Shimodaira-Hasegawa (2002) test.

The high rate of genetic exchange in *X. laevis* could be influenced by large intron size, as seen in salmonid fishes (Shum et al. 2002). We attempted to PCR amplify the intron intervening exons 2 and 3 of the class Ia gene but were unable to recover any specific amplification product. However, bioinformatic exploration of a draft of the *X. tropicalis* genome indicates that a very large intron separates exons 2 and 3 of a class I gene. The largest intron in a class I gene was reported by Shiina et al. (2005); however, the draft genome of *X. tropicalis* indicates that the corresponding intron is much larger (D. H. Bos, unpublished data). If intron size in a closely related species is an indication of class I intron size in *X. laevis*, then it is probable that recombination rates in the class I gene are influenced by these genomic features. The high recombination rate and the location of recombination events support the notion that these events may relieve possible Hill-Robertson interference (Hill and Robertson 1966; Otto and Barton 1997) on the evolution of PBR domains with differing selective pressures (Kaufman et al. 1992).

Phylogenetics of *Xenopus* MHC

Class I alleles fall into lineages that are long lasting and predate certain speciation events within the genus *Xenopus*. These conclusions are upheld when evolutionary relationships are reconstructed with either α1 domain sequences or α2/α3 domain sequences. However, we note that although the sequences may be useful in terms of the test of substitution saturation, our conclusions are based on partitioned data, resulting in short sequences which have less power to resolve and support relationships. In reconstructing evolutionary analysis on MHC genes, other authors have used the "total evidence" approach (Kluge 1989) to establish the species distribution of allelic lineages (e.g., Shum et al. 2001). Such an approach using ML methods on these data also confirm the above conclusions (data not shown).

Based on our reconstruction of evolutionary relationships, transspecies sharing of allelic lineages among certain divergent *Xenopus* species takes place. Comparison of this hypothesis with scenarios that constrain allelic separation

among various species groups confirms this result. Class Ia transspecies evolution in *Xenopus* extends to species thought to be much more evolutionarily divergent than the species among which allele-lineage sharing is commonly found in primate class I genes. While the class Ia sequences from all *Xenopus* species form a monophyletic clade, *X. tropicalis* appears to have diverged from other *Xenopus* species prior to the formation of the extant class Ia lineages. This is not surprising because the divergence time of *X. tropicalis* and the *Xenopus* common ancestor is estimated between 50 and 81 MYA (Evans et al. 2004). Undoubtedly, the transspecies mode of evolution in *Xenopus* is strongly influenced by natural selection acting on the class Ia locus (D. H. Bos and B. Waldman, unpublished data), which tends to extend allele retention.

Pattern of Class Ia MHC Evolution in *X. laevis*

MHC class Ia evolution in *X. laevis* is more similar to MHC evolution in salmonid fishes than mammals. Salmonid fishes and *Xenopus* frogs show similarity in at least three aspects of MHC evolution: (1) levels of polymorphism exceeding that found in primates, (2) a distinct pattern of genetic exchange, and (3) class Ia lineages that are maintained for long periods of time. All three of these characteristics differ from the patterns of class Ia evolution found in primates. For instance, primates exhibit no allele sharing among species thought to last share a common ancestor approximately 35 MYA (Vogel et al. 1999), and recombination events are mostly spread throughout the gene sequence and involve very short sequence tracts (Jakobsen, Wilson, and Easteal 1998).

The pattern of recombination, polymorphism, and transspecies allele sharing of class Ia sequences described in salmonid fishes, and now found in *X. laevis*, may be the result of certain genomic features of the MHC region. Shum et al. (2001) suggested that these features of evolution might be due to the separation of MHC class I and class II regions onto different chromosomes. The nonlinkage of class I and class II regions may influence patterns of MHC evolution by altering the potential Hill-Robertson constraints or selection for conserved haplotype blocks and linkage disequilibrium. However, class Ia evolution in *X. laevis* and salmonid fishes shows similarities in polymorphism, recombination, and transspecies polymorphism. These similarities support the idea that the separation of the class I and class II regions onto different chromosomes is alone insufficient to account for these patterns of MHC evolution.

Other factors may play a role in determining class Ia gene evolution. For instance, both salmonid and *X. laevis* class Ia–processing and class I–processing pathway genes are located close to one another (Namikawa et al. 1995; Takami et al. 1997; Ohta et al. 1999). A likely result of this linkage is that distinctive allelic associations exist among class Ia–processing and class I–processing genes in *Xenopus* frogs and other taxa but are not known in primates and mice (Joly et al. 1998; Kaufman 1999; Ohta et al. 2003). Therefore, it is possible that the pattern of evolution common to *Xenopus* is due in part to the number of class Ia loci, linkage, and possible coevolution of this suite of genes. Additionally, the location of the class Ia gene in a "central" rather than "distal" position within the MHC region may influence patterns of evolution, as suggested by Nonaka et al. (1997).

Acknowledgments

Martin Flajnik and Neil Gemmell provided useful comments on this manuscript, and Louis Du Pasquier provided ideas and invaluable technical assistance in the laboratory and in the animal breeding and care facility. Scott Edwards and two anonymous reviewers provided constructive comments on an earlier version of this manuscript. This research was supported by the Marsden Fund (Royal Society of New Zealand) and a Ph.D. scholarship from the University of Canterbury.

Literature Cited

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control **19**:716–723.

Andolfatto, P., and M. Nordborg. 1998. The effect of gene conversion on intralocus associations. Genetics **148**:1397–1399.

Bingulac-Popovic, J., F. Figueroa, A. Sato, W. S. Talbot, S. L. Johnson, M. Gates, J. H. Postlethwait, and J. Klein. 1997. Mapping of MHC class I and class II regions to different linkage groups in the zebrafish, *Danio rerio*. Immunogenetics **46**: 129–134.

Bontrop, R. E., N. Otting, N. de Groot, and G. G. M. Doxiadis. 1999. Major histocompatibility complex class II polymorphisms in primates. Immunol. Rev. **167**:339–350.

Bos, D. H., and D. Posada. 2005. Using models of nucleotide evolution to build phylogenetic trees. Dev. Comp. Immunol. **29**:211–227.

Consuegra, S., H.-J. Megens, H. Schaschl, K. Leon, R. J. M. Stet, and W. C. Jordan. 2005. Rapid evolution of the MH class I locus results in different allelic compositions in recently diverged populations of Atlantic salmon. Mol. Biol. Evol. **22**:1095–1106.

Edwards, S. V., E. K. Wakeland, and W. K. Potts. 1995. Contrasting histories of avian and mammalian MHC genes revealed by class II B sequences from songbirds. Proc. Natl. Acad. Sci. USA **92**:12200–12204.

Evans, B. J., D. B. Kelley, R. C. Tinsley, D. J. Melnick, and D. C. Cannatella. 2004. A mitochondrial DNA phylogeny of African clawed frogs: phylogeography and implications for polyploid evolution. Mol. Phylogenet. Evol. **33**:197–213.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

Figueroa, F., E. Gunther, and J. Klein. 1988. MHC polymorphism pre-dating speciation. Nature **335**:265–267.

Flajnik, M. F., J. Kaufman, P. Riegert, and L. Du Pasquier. 1984. Identification of class I major histocompatibility complex encoded molecules in the amphibian *Xenopus*. Immunogenetics **20**:134–143.

Flajnik, M. F., Y. Ohta, A. S. Greenberg, L. Salter-Cid, A. Carrizosa, L. Du Pasquier, and M. Kasahara. 1999. Two ancient allelic lineages at the single classical class I locus in the *Xenopus* MHC. J. Immunol. **163**:3826–3833.

Hall, T. 1999. Bioedit: a user-friendly biological sequence alignment editor and analysis program for Window 95/98/NT. Nucleic Acids Symp. Ser. **41**:95–98.

Hess, C. M., and S. V. Edwards. 2002. The evolution of the major histocompatibility complex in birds. Bioscience **52**:423–431.

Hill, W. G., and A. Robertson. 1966. The effect of linkage on limits to artificial selection. Genet. Res. **8**:269–294.

Hughes, A. L., M. K. Hughes, and D. I. Watkins. 1993. Contrasting roles of interallelic recombination at the HLA-A and HLA-B loci. Genetics **133**:669–680.

Hughes, A. L., and M. Yeager. 1998. Natural selection at major histocompatibility complex loci of vertebrates. Ann. Rev. Genet. **32**:415–435.

Jakobsen, I. B., S. R. Wilson, and S. Easteal. 1998. Patterns of reticulate evolution for the classical class I and II HLA loci. Immunogenetics **48**:312–323.

Joly, E., A. F. Le Rolle, A. L. Gonzalez, B. Mehling, J. Stevens, W. J. Coadwell, T. Hunig, J. C. Howard, and G. W. Butcher. 1998. Co-evolution of rat TAP transporters and MHC class I RT1-A molecules. Curr. Biol. **8**:169–172.

Judo, M. S. B., A. B. Wedel, and C. Wilson. 1998. Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res. **26**:1819–1825.

Kaufman, J. 1999. Co-evolving genes in MHC haplotypes: the "rule" for nonmammalian vertebrates? Immunogenetics **50**:228–236.

Kaufman, J., R. Anderson, D. Avila, J. Engberg, J. Lambris, J. Salomonsen, K. Welinder, and K. Skjodt. 1992. Different features of the MHC class I heterodimer have evolved at different rates. J. Immunol. **142**:1532–1546.

Kelley, L., J. Walter, and J. Trowsdale. 2005. Comparative genomics of major histocompatibility complexes. Immunogenetics **56**:683–695.

Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78**:454–458.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicurates* (Boidae, Serpentes). Syst. Zool. **38**:7–25.

Lawlor, D. A., F. E. Ward, P. D. Ennis, A. P. Jackson, and P. Parham. 1988. *HLA-A* and *HLA-B* polymorphism predate the divergence of humans and chimpanzees. Nature **335**:268–271.

Martin, D. P., C. Williamson, and D. Posada. 2005. RDP2: recombination detection and analysis from sequence alignments. Bioinformatics **21**:260–262.

McVean, G., P. Awadalla, and P. Fearnhead. 2002. A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160**:1231–1241.

Namikawa, C., L. Salter-Cid, M. F. Flajnik, Y. Kato, M. Nonaka, and M. Sasaki. 1995. Isolation of Xenopus *LMP-7* homologues: striking allelic diversity and linkage to MHC. J. Immunol. **155**:1964–1971.

Nonaka, M., C. Namikawa, Y. Kato, M. Sasaki, L. Salter-Cid, and M. F. Flajnik. 1997. Major histocompatibility complex gene mapping in the amphibian *Xenopus* implies a primordial organization. Proc. Natl. Acad. Sci. USA **94**:5789–5791.

Ohta, Y., K. Okamura, E. C. McKinney, S. Bartl, K. Hashimoto, and M. F. Flajnik. 2000. Primitive synteny of vertebrate histocompatibility complex class I and class II genes. Proc. Natl. Acad. Sci. USA **97**:4712–4717.

Ohta, Y., S. J. Powis, W. J. Coadwell, D. E. Haliniewski, Y. Liu, H. Li, and M. F. Flajnik. 1999. Identification and mapping of *Xenopus* TAP2 genes. Immunogenetics **49**:171–182.

Ohta, Y., S. J. Powis, R. L. Lohr, M. Nonaka, L. Du Pasquier, and M. F. Flajnik. 2003. Two highly divergent ancient allelic lineages of the transporter associated with antigen processing (TAP) gene in *Xenopus*: further evidence for co-evolution among MHC class I region genes. Eur. J. Immunol. **33**:3017–3027.

Otto, S. P., and N. H. Barton. 1997. The evolution of recombination: removing the limits to natural selection. Genetics **147**:879–906.

Parham, P., and T. Ohta. 1996. Population biology of antigen presentation by MHC class I molecules. Science **272**:67–74.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14**:817–818.

———. 2001*a*. Intraspecific gene genealogies: trees grafting into networks. Trends Ecol. Evol. **16**:37–45.

———. 2001*b*. Selecting models of nucleotide substitution: an application to human immunodeficiency virus (HIV-1). Mol. Biol. Evol. **18**:897–906.

———. 2001*c*. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. USA **98**:13757–13762.

Rozen, S., and H. Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Pp. 365–386 *in* S. Krawetz and S. Misener, eds. Bioinformatics methods and protocols: methods in molecular biology. Humana Press, Totowa, N.J.

Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Sambrook, E., F. Fritsch, and T. Maniatis. 1989. Molecular cloning. Cold Spring Harbor Press, Cold Spring, N.Y.

Shiina, T., J. M. Dijkstra, S. Shimizu et al. (15 co-authors). 2005. Interchromosomal duplication of major histocompatibility complex class I regions in rainbow trout (*Oncorhynchus mykiss*), a species with a presumably recent tetraploid ancestry. Immunogenetics **56**:878–893.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. **51**:492–508.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114–1116.

———. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics **17**:1246–1247.

Shum, B. P., D. Avila, L. Du Pasquier, M. Kasahara, and M. F. Flajnik. 1993. Isolation of a classical MHC class I cDNA from an amphibian. Evidence for only one class I locus in the *Xenopus* MHC. J. Immunol. **151**:5376–5386.

Shum, B. P., L. A. Guethlein, L. R. Flodin, M. A. Adkinson, R. P. Hedrick, R. B. Nehring, R. J. M. Stet, C. Secombes, and P. Parham. 2001. Modes of salmon MHC class I and II evolution differ from the primate paradigm. J. Immunol. **166**: 3297–3308.

Shum, B. P., P. M. Mason, K. E. Magor, L. R. Flodin, R. J. M. Stet, and P. Parham. 2002. Structures of two major histocompatibility complex class I genes of the rainbow trout (*Oncorhynchus mykiss*). Immunogenetics **54**:193–199.

Swofford, D. L. 1998. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4.0. Sinauer, Sunderland, Mass.

Takami, K., Z. Zaleska-Rutczynska, F. Figueroa, and J. Klein. 1997. Linkages of *LMP*, *TAP*, and *RING3* with *Mhc* class I rather than class II genes in the zebrafish. J. Immunol. **159**:6052–6060.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

Vogel, T. U., D. T. Evans, J. A. Urvater, D. H. O'Connor, A. L. Hughes, and D. I. Watkins. 1999. Major histocompatibility complex class I genes in primates: co-evolution with pathogens. Immunol. Rev. **167**:327–337.

Xia, X., Z. Xie, M. Salemi, L. Chen, and Y. Wang. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. **26**:1–7.

Scott Edwards, Associate Editor

Accepted August 31, 2005